L'organisation des connaissances au prisme du langage, du texte et du discours. Un parcours en recherche d'information.

Viviane Clavier

UNIVERSITE STENDHAL – GRENOBLE 3
Groupe de Recherche sur les Enjeux de la Communication

L'organisation des connaissances au prisme du langage, du texte et du discours.

Un parcours en recherche d'information

Mémoire pour l'obtention de l'habilitation à diriger des recherches en sciences de l'information et de la communication

Viviane Clavier

sous le parrainage de Geneviève Lallich-Boidin et Isabelle Pailliart

Volume 1

HDR soutenue publiquement le 10 mars 2014 à l'Institut de la Communication et des Médias devant le jury composé de :

Laurence Balicco,
Professeure en sciences de l'information et de la communication
Université Pierre Mendès-France, Grenoble 2. Examinatrice

Stéphane Chaudiron,
Professeur en sciences de l'information et de la communication
Université Charles de Gaulle, Lille 3. Rapporteur

Viviane Couzinet,
Professeure en sciences de l'information et de la communication
Université Paul Sabatier, Toulouse 3. Présidente du jury

Francis Grossmann,
Professeur en sciences du langage
Université Stendhal, Grenoble 3. Rapporteur

Geneviève Lallich-Boidin,
Professeure émérite en sciences de l'information et de la communication – Université Claude Bernard, Lyon 1. Co-directrice

Isabelle Pailliart,
Professeure en sciences de l'information et de la communication
Université Stendhal, Grenoble. Co-directrice

Remerciements

Je remercie très chaleureusement Geneviève Lallich-Boidin et Isabelle Pailliart qui m'ont parrainée dans la réalisation de ce mémoire d'habilitation à diriger des recherches. Je suis heureuse que Geneviève Lallich-Boidin, après avoir co-dirigé mon mémoire de maîtrise, de DEA et de thèse, soit encore à mes côtés : son regard critique est toujours aussi stimulant. Je suis très reconnaissante à Isabelle Pailliart pour la rigueur scientifique, la finesse d'analyse et la pertinence de ses relectures. Sans la confiance qu'elle m'a témoignée, je n'aurais jamais réalisé ce travail.

J'adresse mes remerciements les plus sincères aux membres du jury qui ont accepté d'évaluer mes travaux et d'assister à la soutenance : Stéphane Chaudiron, Viviane Couzinet et Francis Grossmann. Je suis très honorée de leur présence et leur sais gré de la lecture attentive qu'ils ont faite de mon mémoire. Merci à Laurence Balicco avec qui je partage une culture commune, et, le temps faisant, l'histoire d'une équipe.

Je remercie Fabienne Martin-Juchat, directrice du département des sciences de l'information et de la communication, qui m'a officiellement dispensée d'assurer la responsabilité de filières afin de me consacrer à la rédaction de ce mémoire.

Je remercie de tout cœur mes collègues de travail, qui, par leurs encouragements et leur qualité de chercheurs m'ont aidée à construire ma pensée. Leur présence dans ce laboratoire est très stimulante. Je pense tout particulièrement à Céline Paganelli avec qui je mène des recherches depuis de nombreuses années, mais aussi à Benoit Lafon, pour ses conseils avisés, Evelyne Mounier et Adrian Staii, pour leur soutien. Merci également à Agnès Tutin et Françoise Boch qui m'ont réservé un accueil chaleureux au Lidilem et à Guillaume Cleuziou qui a accepté de nombreuses collaborations.

Je remercie enfin Josiane Clavier pour ses relectures, l'intérêt qu'elle porte à mes travaux ; mes amies, en particulier Christine Peponnet, pour son ouverture d'esprit et ses compétences en linguistique anglaise ; mes enfants, Lucas et Anne Méthivier, pour leur bonne humeur.

Table des matières

INTRODUCTION **4**

PREMIERE PARTIE. ORGANISATION DES CONNAISSANCES ET SYSTEMES DE RECHERCHE D'INFORMATION : LA PLACE DU LANGAGE **13**

1. LE LANGAGE COMME MODELE D'ORGANISATION DES CONNAISSANCES : *NORMALISER, INTERPRETER ET DECOMPTER* **18**

1.1. L'environnement de notre recherche : éléments de cadrage **18**
1.1.1. Le TAL et la recherche d'information 18
1.1.2. La morphologie dérivationnelle et la recherche d'information 20
1.1.3. Le CRISS : un cadre de recherche pluridisciplinaire 22
1.2. L'apport de la linguistique à l'indexation automatique **25**
1.2.1. Nos choix d'indexation 27
1.2.2. Nos choix théoriques de modélisation pour le TAL 38
1.2.3. Organisation des connaissances dans un environnement distribué 45
1.3. Bilan critique **63**
1.3.1. Les apports à l'organisation des connaissances 63
1.3.2. Les limites 67
1.3.3. Pour conclure sur l'apport de la morphologie à la recherche d'information 70

2. LES TEXTES COMME MODELE D'ORGANISATION DES CONNAISSANCES : *PROFILER, ANNOTER, CLASSER* **71**

2.1. L'environnement de notre recherche : éléments de cadrage **72**
2.1.1. La classification automatique et la recherche d'information 72
2.1.2. La linguistique de corpus et le TAL 73
2.1.3. Le LIFO et le CORAL : des cadres de recherche mono-disciplinaire 74
2.2. L'apport de la linguistique à la classification **77**
2.2.1. Classification, distance mathématique et données textuelles 77
2.2.2. Caractérisation du genre textuel et linguistique de corpus 84
2.2.3. Applications à la recherche d'information 95
2.3. Bilan critique **101**
2.3.1. Les apports à l'organisation des connaissances 101
2.3.2. Les limites 106
2.3.3. Pour conclure sur l'organisation des connaissances 109

DEUXIEME PARTIE. ORGANISATION DES CONNAISSANCES ET DISPOSITIFS INFORMATIONNELS : VERS UNE MISE EN CONTEXTE **111**

1. L'ENVIRONNEMENT DE NOTRE RECHERCHE : ELEMENTS DE CADRAGE **115**

1.1. Le GRESEC : un cadre de recherche en SIC **115**
1.1.1. Les sciences de l'information se dégagent de l'ingénierie 115
1.1.2. L'enseignement des sciences de l'information 118
1.1.3. Notre participation à des recherches finalisées 120
1.2.4. Notre contribution à un programme de recherche dans l'axe C.I.D 124

1.2. Le contexte dans les courants de la recherche d'information **128**
1.2.1. Le contexte pour les approches sociales 129
1.2.2. Le contexte pour la linguistique 131
1.2.3. Le contexte pour l'informatique 133
1.2.4. Pour conclure sur « les contextes » 135

2. DES METHODES POUR ORGANISER L'INFORMATION : LE ROLE DES DISCOURS **136**

2.1. La liste des corpus analysés **137**
2.2. Les entretiens et les protocoles verbaux **139**
2.3. L'analyse des discours **140**
2.3.1. La finalité des analyses de corpus 140
2.3.2. Trois conceptions de corpus 144
2.3.3. Analyser le contexte : un construit méthodologique 153
2.3.4. Conclusion sur la place de la méthodologie dans notre recherche 159

3. LES DISCOURS POUR GUIDER L'ORGANISATION DES CONNAISSANCES : *LIRE, ANNOTER, COMMUNIQUER…* **161**

3.1. L'évaluation n'est pas le seul enjeu de l'organisation des connaissances **161**
3.2. Il existe deux visions possibles de l'organisation des connaissances **163**
3.3. Différents éléments influencent l'organisation des connaissances **166**
3.3.1. NOESIS : la terminologie médicale, les schémas et les articles de revue 167
3.3.2. Métilde : les variantes formelles, le support papier et le livre 169
3.3.3. Bilan : des langages hiérarchisés aux marques formelles spatialisées 171
3.4. L'analyse des discours contribue à faire émerger des connaissances **173**
3.4.1. CaNu XIX : l'indexation de thèmes s'appuie sur les événements 174
3.4.2. SCIENTEXT : les thèses de doctorat créent un « espace de sens qui favorise la recherche d'un *positionnement* scientifique » 183
3.5. Bilan critique **194**

CONCLUSION **200**

BIBLIOGRAPHIE **205**

ANNEXE 1 : EXTRAIT DU CORPUS CFM BALISE SUIVANT LA NORME TEI **232**

ANNEXE 2 : EXTRAIT DU CORPUS CFM ETIQUETE ET POST-EDITE **236**

ANNEXE 3 : GRAPHE DE COOCCURRENCES DES VALEURS FIGUREES DU MOT *CANCER* DANS LES CORPUS *FRANTEXT*, *LE FIGARO*, ET LE *NOUVEL OBS*. **237**

ANNEXE 4 : GRAPHE DE COOCCURRENCE DES VALEURS PROPRES DU MOT *CANCER* DANS LES CORPUS *FRANTEXT ET LE NOUVELOBS*. **238**

Introduction

Ce mémoire d'habilitation à diriger des recherches présente une synthèse des travaux réalisés seule ou en collaboration depuis la soutenance de notre thèse de doctorat en 1996, soit 17 ans consacrés à l'enseignement et la recherche en sciences de l'information et de la communication (SIC). Organisé en deux volumes, ce document comporte un bilan critique de notre parcours scientifique et une mise en perspective épistémologique, théorique et méthodologique de nos objets de recherche. Le second volume contient une sélection de nos publications classées par ordre chronologique croissant, cet ordre de présentation figurant également dans le premier volume.

Le fil directeur de notre activité de recherche est lié à la description et la caractérisation de contenus textuels, à leur représentation sous la forme de connaissances organisées et médiatisées par des dispositifs pour la recherche d'information.

Le langage tient dans notre recherche une place privilégiée. Nous l'avons appréhendé dans son rapport à la langue comme système articulé en « niveaux » calqués sur les unités de langue, telles la morphologie, la syntaxe et la sémantique. Nous l'avons également analysé en recourant à des méthodes descriptives, convaincue que la modélisation pour le traitement automatique des langues ne pouvait faire l'économie d'une description fine des phénomènes linguistiques et des modèles formels. Nous l'avons saisi à différents paliers de la textualité, le texte étant un lieu d'écriture normée où se manifestent la structure hiérarchique, la thématique et le métadiscours. Le texte a également été considéré dans sa relation au document dont il constitue la face sémiotique, et comme espace discursif où se tissent des relations inter- et intratextuelles. Le langage est pensé comme composante essentielle de l'organisation des connaissances, et, de ce point de vue il est travaillé dans une perspective appliquée : la recherche d'information.

L'organisation des connaissances désigne, au sens étroit, la confection de produits d'information destinés à représenter, organiser et structurer des connaissances pour améliorer la recherche d'information. Yolla Polity, Gérard Henneron et Rosalba Palermiti énumèrent ces produits, appelés également systèmes d'organisation des connaissances (SOC)[1] :

> *« Par organisation des connaissances, il faut entendre :*
>
> *Toutes sortes de schémas d'organisation allant des simples listes alphabétiques ou faiblement structurées (listes d'autorité, glossaires, dictionnaires, nomenclatures, etc.) à des schémas classificatoires hiérarchiques (plans de classement, classifications générales ou spécialisées, taxinomies, listes de vedettes matières, etc.) ou encore à des organisations privilégiant des relations non exclusivement hiérarchiques (thésaurus, réseaux sémantiques, ontologies, etc.), portant sur toutes sortes d'objets allant des documents au sens classique du terme (textes, images fixes et animées, enregistrements sonores, etc.) jusqu'à l'ensemble des phénomènes concrets ou abstraits que l'on peut avoir besoin de recenser, d'organiser et de traiter (objets, événements, processus, etc.) et avec des buts et des objectifs divers : retrouver, enseigner, produire de nouvelles connaissances, communiquer, appliquer des traitements appropriés, etc. » (Polity et al. , 2005 : p. 13)*

Au sens large, l'organisation des connaissances est un champ de recherche qui bénéficie d'une visibilité internationale forte, surtout dans le courant de la Library of Information Science (LIS) où ce thème fait l'objet d'enseignements spécifiques. C'est ainsi que Birger Hjørland, l'un des spécialistes renommé internationalement en sciences de l'information, est professeur en « organisation des connaissances » à la *Royal School of Library and Information Science* à Copenhague. Il existe en outre une revue internationale consacrée à ce thème (*Knowledge Organization*), adossée à une société savante, *International Society for Knowledge Organization*, (ISKO) qui organise des conférences tous les deux ans au niveau international ainsi que des « chapitres-ISKO » qui se déclinent dans différents pays sur des thématiques particulières. Ces congrès et revues sont des lieux d'échanges et de visibilité très importants pour les chercheurs des sciences de l'information, qui, contrairement aux chercheurs des sciences de la communication, bénéficient d'un faible nombre de revues au sein des SIC. Entre ces deux conceptions de l'organisation des connaissances, l'une très pragmatique et l'autre épistémologique, figure l'organisation des connaissances

[1] en anglais *Knowledge Organization Systems (KOS).*

comme *« objet d'étude des processus cognitifs et des techniques intellectuelles qui permettent de classer, indexer, formaliser et modéliser le réel »* (Polity et *al.* 2005, p.13). C'est précisément dans cette perspective que nous posons la question de l'organisation des connaissances, la recherche d'information étant alors un domaine structurant pour appréhender notre objet, pour développer nos méthodes et pour installer notre cadre théorique. La figure suivante présente, à titre d'illustration, les principaux concepts mobilisés dans le mémoire ainsi que leur articulation. Elle distingue plusieurs plans d'analyse qui se répartissent entre une partie gauche dédiée au langage et une partie droite, au contexte social. L'analyse sépare de manière artificielle les dimensions sociale et linguistique et notre travail consiste à les retravailler de manière dynamique en mobilisant des concepts qui se répartissent à leur tour sur trois niveaux : celui du document, celui de la textualité et celui des connaissances. La flèche centrale représente la dynamique de construction et d'organisation des connaissances. Ce processus fait intervenir des outils, des méthodes d'analyse et de traitement de l'information qui s'appliquent aux données. Ces dernières sont appréhendées au prisme du langage, du texte et du discours. La technique est évacuée de cette description, nous l'introduisons ci-après.

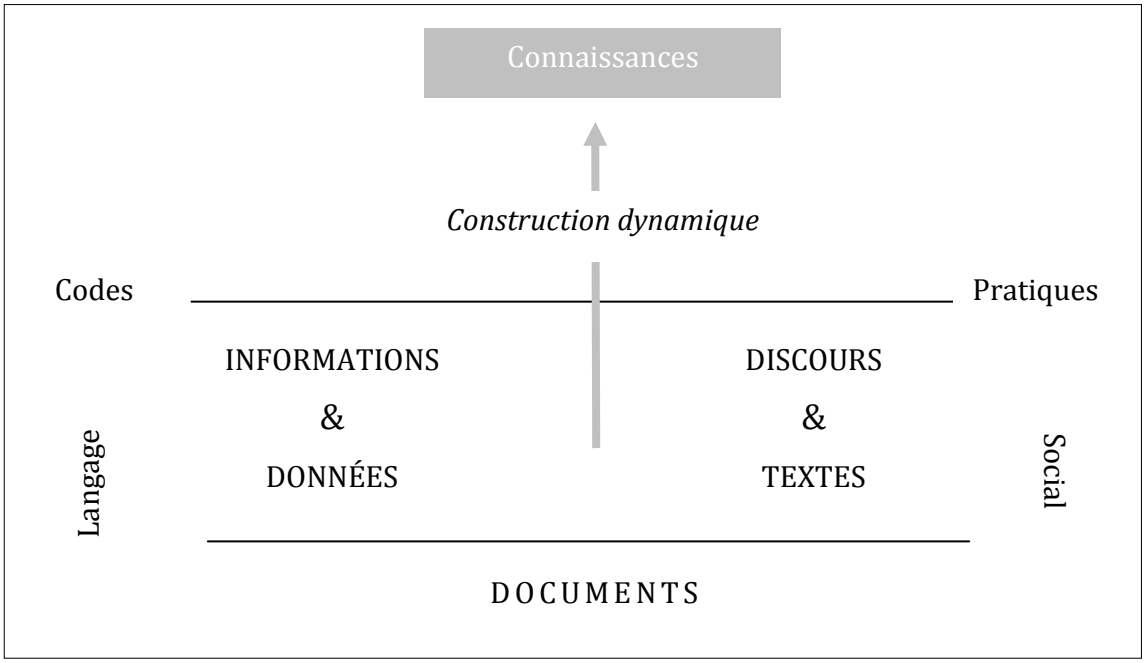


Figure. Positionnement des concepts

L'organisation des connaissances pour la recherche d'information a été envisagée suivant deux grandes orientations qui révèlent une évolution de notre positionnement scientifique et qui correspondent à deux ensembles de travaux présentés dans des parties distinctes.

Dans le premier ensemble[2], l'objectif est de définir, de structurer et de représenter des connaissances linguistiques destinées à être intégrées dans des systèmes de recherche d'information ou à être utilisées dans des méthodes de traitement de l'information. Ce contexte appliqué nous a conduit à travailler les textes dans l'optique d'améliorer différentes méthodes et applications : l'indexation automatique, l'extraction de connaissances, l'analyse automatique de discours, la classification de textes et de mots. En termes de niveaux linguistiques, nous avons plus spécifiquement travaillé la morphologie dérivationnelle, mais nous avons également fait des incursions dans d'autres domaines de la linguistique, étant donné le caractère transversal de la morphologie. Quels que soient les niveaux considérés (morphologie, syntaxe, sémantique) et les méthodes utilisées (TAL, classification), les connaissances ont toujours été appréhendées dans l'optique de favoriser l'appariement entre des requêtes et des documents. La conception de la recherche d'information qui préside à cette première série de travaux a été envisagée dans des contextes fortement marqués par les enjeux scientifiques anglo-saxons de *l'information retrieval*, i.e une approche qui privilégie les outils, les techniques et les méthodes destinées au développement des systèmes de recherche d'information (SRI).

Dans le second ensemble de travaux[3], nous avons appréhendé les dispositifs dans leur contexte social, cette posture transparaissant dans les dénominations qui ont cours dans la discipline : « dispositifs d'accès à l'information », « dispositifs informationnels », ou « dispositifs info-communicationnels ». Cette conception de la recherche d'information est connue dans la littérature anglo-saxonne sous diverses appellations : *information seeking, information searching, information behaviour, etc.*

[2] Les références des publications correspondant à ce premier ensemble sont numérotées de 1 à 29 (cf. volume 2, liste des publications classées par ordre chronologique)

[3] Les références des publications correspondant à ce deuxième ensemble sont numérotées de 30 à 45 (cf. vol. 2, liste des publications classées par ordre chronologique)

Bien que ces dénominations ne soient pas équivalentes, toutes ont en commun de décrire l'activité de recherche d'information en contexte (de travail, de loisirs) et de décrire les comportements humains de populations (des professionnels, des séniors, des enfants, etc.). Ce glissement théorique induit un changement de posture pour appréhender l'organisation des connaissances : l'objectif consiste alors à relier les dimensions qui émergent d'une part des usages et des pratiques informationnelles, et d'autre part, des propriétés linguistiques de l'information, perpétuant ainsi la tradition documentaire qui définit ses traitements en fonction des contenus des documents et de leurs publics. Elargissant notre réflexion jusque là guidée par la conception de SRI, nous avons participé à des études sur les pratiques informationnelles d'usagers en situation de travail. Ce faisant, nous observions une grande diversité d'activités informationnelles et communicationnelles comme la lecture, la navigation, l'écriture, l'annotation critique, le partage d'information, etc. Ces activités qui révèlent un rapport singulier à l'information constituent le point de départ de notre contribution à l'organisation des connaissances.

L'organisation des connaissances est interrogée d'abord sous l'angle technique, puis social et pose la question de la médiation, un champ de recherche important en SIC. Pour l'information-documentation, la médiation documentaire fait traditionnellement référence aux « activités des professionnels de l'information qui s'appuient sur le traitement documentaire » (Liquète et al., 2010). Dans un article récent portant sur la médiation numérique documentaire, Cécile Gardiès et Isabelle Fabre précisent cette notion en mettant en perspective des dimensions théoriques (l'information et la communication), techniques et organisationnelles (des agencements de matériels) et enfin une dimension humaine (qui tient compte du social)

> *« La médiation documentaire concerne une médiation des savoirs mettant en place, grâce à un tiers, des interfaces qui accompagnent l'usager et facilitent les usages. Elle permet de concilier deux choses jusque-là non rassemblées pour établir une communication et un accès à l'information. C'est par sa capacité à lier information et communication qu'elle peut être qualifiée de médiation documentaire. Elle s'appuie sur des composants humains ou matériels qu'on peut distinguer en « médiateurs sociaux « naturels » (normes, valeurs…), médiateurs humains (négociateurs, chefs…), dispositifs complexes (agencements matériels et géographiques, organisationnels et techniques…) » La médiation est donc fortement liée à la*

> *question du dispositif étudié en SIC comme objet matériel médiateur qui désigne « l'ensemble des substrats matériels de la communication » (Gardiès et Fabre, 2012)*

De manière proche, nous utilisons le terme de médiation dans le sens de « traduction », de « connexion » ou de « lien » (Liquète, et al. 2010) entre des éléments de différentes natures. Cependant, notre focale est plus restreinte puisque nous nous intéressons à un aspect seulement des dispositifs : les connaissances. Dans la première série de nos travaux, l'enjeu était la conception de SRI. La médiation était uniquement technique, puisque, pour choisir les connaissances, les organiser et les modéliser nous nous appuyions sur les composantes internes des SRI : la langue vue sous l'angle de l'ordinateur, les textes fournis en entrée du SRI, les grammaires, les dictionnaires qui participent de la conception d'un SRI. Dans la seconde série de travaux en revanche, l'enjeu n'est plus prioritairement la conception. La question des connaissances se pose à plusieurs niveaux : celui des pratiques informationnelles, des usages des « dispositifs médiateurs », des contenus, des documents, etc. Cette perspective socio-technique pose la question de la complexité méthodologique et théorique pour recueillir, caractériser et organiser les connaissances.

Nous avons fait le choix de présenter nos travaux de recherche de manière chronologique et nous montrons le lien entre l'évolution de notre positionnement scientifique et les quatre contextes institutionnels que nous avons connus :

1) à Grenoble, le Centre de Recherche en Informatique appliquée aux Sciences Sociales (CRISS) à l'Université Pierre Mendès-France(1990-1996) ;
2) à Orléans, le Laboratoire d'Informatique Fondamentale d'Orléans à (LIFO) à l'Université d'Orléans (1998-2004) ;
3) à Orléans, le Centre Orléanais de Recherche en Anthropologie et Linguistique (CORAL) à l'Université d'Orléans (2001-2004) ;
4) à Grenoble, le Groupe de Recherche Sur les Enjeux de la Communication (GRESEC) à l'Université Stendhal (depuis 2004).

Le premier laboratoire de recherche que nous avons connu est le Centre de Recherches en Informatique appliquée aux Sciences Sociales (CRISS, Université Pierre

Mendès-France, Grenoble2) dirigé par Jacques Rouault, professeur en sciences de l'information et de la communication. Nous y avons préparé notre thèse de doctorat[4]. Ce laboratoire a connu une existence d'une dizaine d'années à l'Université Pierre Mendès-France, puis en 1993, une partie des membres a été intégrée au GRESEC à l'Université Stendhal. Les travaux menés au CRISS sous la direction de Jacques Rouault relevaient de l'informatique documentaire. Mathématicien de formation, Jacques Rouault a suivi les cours d'Antoine Culioli, linguiste de renom, connu pour avoir développé une théorie de l'énonciation. A Grenoble, Jacques Rouault a préparé une thèse de doctorat sous la direction de Bernard Vauquois[5], le père de la traduction automatique en France. Bernard Vauquois avait impulsé localement une dynamique importante autour du traitement automatique des langues. Il régnait dans les différentes équipes grenobloises une très forte adhésion au paradigme de *l'Information Science* qui émanait des Etats-Unis.

Les deuxième et troisième laboratoires remontent à la période 1998-2004, où nous avons obtenu notre premier poste de maître de conférences en sciences de l'information et de la communication à l'IUT d'Orléans. Il n'y avait pas de laboratoire de recherche dans notre discipline. Nous avons d'abord été accueillie comme membre permanent au Laboratoire d'Informatique Fondamentale d'Orléans (LIFO) qui comptait une équipe de chercheurs travaillant sur le langage naturel et la recherche d'information sous la direction de Jean-Claude Bassano, professeur d'informatique. Puis, à partir de 2001, le laboratoire, ayant obtenu une association au CNRS en tant que Formation de Recherche en Evolution, a été restructuré. Nous sommes alors devenue membre associé au LIFO, dans l'équipe Contraintes et Apprentissages sous la direction de Christel Vrain, professeur d'informatique. Parallèlement, nous avons été accueillie comme membre permanent au Centre Orléanais de Recherche en Anthropologie et Linguistique (CORAL) sous la direction de Gabriel Bergounioux,

4 Clavier Viviane, *Modélisation de la suffixation pour le traitement automatique du français. Application à la recherche d'information,* thèse de doctorat en sciences de l'information et de la communication soutenue en 1996, sous la direction de Jaques Rouault, Université Stendhal, Grenoble 3.

5 Rouault Jacques, *Approche formelle de problèmes liés à la sémantique des langues naturelles*, thèse de doctorat en mathématiques soutenue en 1971, sous la direction de Bernard Vauquois, Université Joseph Fourrier.

professeur de linguistique. Les axes de recherche du CORAL portaient notamment sur la description des langues, et en particuliers des « langues rares », comme les créoles et le palikur.

Le quatrième contexte institutionnel est le Groupe de Recherche sur les Enjeux de la Communication (GRESEC) à l'Université Stendhal de Grenoble 3 spécialisé dans les recherches en sciences de la communication et dirigé actuellement par Isabelle Pailliart, professeure en sciences de l'information et de la communication. Nous avons intégré ce laboratoire en 2004, à la faveur d'une mutation. Le Gresec est structuré en quatre axes, nous participons à l'axe Connaissance, Recherche d'information, Interfaces et Systèmes de Traitement Automatique de la Langue (CRISTAL), rebaptisé en 2011 Connaissance Information Document (C.I.D).

Lors de notre expérience de recherche, nous avons pris part à plusieurs contextes théoriques et disciplinaires qui abordaient la recherche d'information tantôt suivant une perspective technique, tantôt humaine et sociale, même s'il n'existe pas d'approches exclusives l'une de l'autre. Ces différents ancrages ont eu une influence décisive sur notre conception de l'organisation des connaissances : sur les formes des représentations et leurs significations, sur les méthodologies de recueil des connaissances, sur la modélisation, voire, sur la formalisation des connaissances**.**

La première partie de ce mémoire aborde l'organisation des connaissances dans la perspective appliquée de développement de systèmes de recherche d'information.

Cette partie comporte deux chapitres. Le premier chapitre s'intéresse aux connaissances qui s'appliquent aux unités de rang inférieur aux mots, les morphèmes. Ces unités sont décrites au moyen d'informations qui relèvent de différents niveaux linguistiques : morphophonologie, syntaxe et sémantique. Les connaissances sont définies pour être intégrées dans un analyseur morphologique du français destiné à reconnaître des mots construits par suffixation. Deux applications sont destinées à valider le modèle morphologique : l'extraction de connaissances dans un manuel d'utilisation et la contribution au développement d'un prototype d'analyse

automatique du discours. Le second chapitre montre comment la prise en compte de connaissances linguistiques permet d'améliorer les méthodes de classification pour la recherche d'information. Ces connaissances utilisent des informations linguistiques propres à caractériser le genre textuel et sont annotées dans des documents également normés et standardisés : les documents XML annotés au moyen de la norme TEI[6]. Une évaluation des méthodes de classification est réalisée pour apprécier l'apport respectif des connaissances morphosyntaxiques *versus* lexicales pour classer des textes en genres textuels et en thèmes.

La seconde partie aborde l'organisation des connaissances en lien avec l'observation des pratiques sociales des dispositifs informationnels.

Cette partie comporte trois chapitres. Le premier chapitre dresse notamment un état des lieux de la question des contextes en recherche d'information pour les sciences sociales, la linguistique et l'informatique. Elle met en lumière le changement de posture théorique et méthodologique pour aborder la recherche d'information dans une perspective sociale et interdisciplinaire. Le deuxième chapitre présente des aspects méthodologiques liés à la constitution de corpus pour recueillir et analyser des discours destinés à organiser et structurer l'information. Le troisième montre l'apport et les limites de nos méthodes pour l'organisation des connaissances.

[6] Text Encoding Initiative.

Première partie. Organisation des connaissances et systèmes de recherche d'information : la place du langage

Dans cette partie, nous montrons les apports de la linguistique pour définir, structurer, organiser et représenter des connaissances qui interviennent dans des applications pour la recherche d'information : l'extraction de connaissances, l'analyse automatique de discours et la veille. Les traitements de l'information qui font intervenir ces connaissances linguistiques sont l'indexation en texte intégral et la classification automatique.

Nous abordons le traitement et la modélisation du langage dans la perspective du courant anglo-saxon de *l'information retrieval* qui privilégie les outils, les techniques et les méthodes destinés au développement des systèmes de recherche d'information (SRI). Ce courant remonte aux années 50. Il marque les grands débuts des travaux de la science de l'information (*Information Science*) et de l'informatique (*Computer Science*) et est concomitant du développement des ordinateurs susceptibles d'accomplir des tâches dites « automatisables »[7]. Ainsi, la recherche d'information et la traduction automatique[8] étaient-elles appliquées à l'information scientifique et technique dans un cadre militaire. Pour Calvin N. Moeers, la recherche d'information consiste à trouver une information dont l'existence et la localisation sont *a priori* inconnues[9].

La plupart des travaux issus du courant de *l'information retrieval* reposent sur des objectifs annoncés dès la fin des années 70 par Cyril Cleverdon, Wilfrid Lancaster, Gerard Salton, Peter Luhn, etc. Dans l'introduction de son ouvrage intitulé *Information Retrieval*, van Rijsbergen (1979) mentionne que la recherche d'information est envisagée comme un processus automatisé consistant à récupérer

[7] Von Neumann publie en 1946 des propositions théoriques fondamentales sur le fonctionnement d'un ordinateur, le premier calculateur *The Electroniqic Delay Storage Automatic Calculator* (EDSAC) est créé en 1946, les principes de la « récupération d'information » sont énoncés dans la foulée (Ollivier, 2007 : p. 176-177).

[8] Les premiers projets de traduction automatique remontent aux années 40 et se sont développés dans le contexte de la guerre froide, notamment après le lancement par les soviétiques du Spoutnik en 1957 (Cori, 2008).

[9] « The requirements of information retrieval, of finding information whose location or very existence is a priori unknown. » (Garfield, 1997)

des documents – l'information étant le substitut du document – sans pour autant que ce système prétende « informer », au sens de « changer la connaissance »[10]. Le système permet simplement de renseigner sur l'existence ou la non-existence d'une information, et permet de fournir la référence du « document » demandé. Il existe dès les débuts un flou sur la nature de l'objet recherché, l'information étant assimilée au document, préoccupation qui a marqué l'âge d'or de *l'informatique documentaire*, appelée encore *documentation automatique*. Avec le développement de l'internet, la disponibilité des ressources de différents statuts sur le web conduit à une certaine indifférenciation des types de recherche : la *recherche d'information* englobe ainsi un ensemble « d'objets » plus ou moins opaques pour l'usager.[11]

Quel que soit l'objet recherché, le principe d'un système de recherche d'information est toujours le même : un système de recherche d'information fournit un ensemble de documents « pertinents » en réponse à une demande (Lallich-Boidin et Maret, 2005 : p.11). Geneviève Lallich-Boidin et Dominique Maret précisent qu'un système de recherche d'information ne compare pas directement les documents aux demandes : l'appariement s'opère sur des représentations qui résultent de transformations prenant en compte le contexte du document et de la requête. Les auteurs évoquent la diversité des produits issus de ces représentations : les index produits

[10] « *Information retrieval is a wide, often loosely-defined term but in these pages I shall be concerned only with automatic information retrieval systems. Automatic as opposed to manual and information as opposed to data or fact. Unfortunately the word information can be very misleading. In the context of information retrieval (IR), information, in the technical meaning given in Shannon's theory of communication, is not readily measured […]. In fact, in many cases one can adequately describe the kind of retrieval by simply substituting 'document' for 'information'. Nevertheless, 'information retrieval' has become accepted as a description of the kind of work published by Cleverdon, Salton, Sparck Jones, Lancaster and others. A perfectly straightforward definition along these lines is given by Lancaster […] : 'Information retrieval is the term conventionally, though somewhat inaccurately, applied to the type of activity discussed in this volume.* ***An information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request.*** **»** (van Rijsbergen, 1979) – c'est nous qui soulignons.

[11] Suivant Jacques Maniez, « *le rapprochement entre la recherche d'information et la recherche d'objets de toute nature n'a guère sollicitée la curiosité de spécialistes de l'information. Mon hypothèse est qu'on peut tirer de cette comparaison quelques enseignements sur la place du langage dans toutes les techniques de « retrouvage » et, à partir de là, sur la spécificité de la recherche documentaire. Dans cette perspective, on considère l'information comme l'ensemble des informations – ou objets informationnels - et les systèmes de recherche d'information (SRI) comme un sous-ensemble d'une catégorie plus générale que nous désignons, faute de mieux, par « systèmes de recherche d'objets » ou « systèmes de fourniture d'objets à la demande* » […] » (Maniez, 1994 : p. 15)

automatiquement par les moteurs de recherche sur du texte intégral, toutes les métadonnées, y compris celles qui résultent du catalogage comme les données bibliographiques. Depuis que les bibliothèques existent, sont produites des représentations : registres d'inventaires, notices bibliographiques, catalogues, etc. (*ibid*., p. 14). Ainsi, suivant cette conception, les *représentations* et les *connaissances* sont-elles assimilées, toutes deux contribuant à des formes de médiation documentaire, la première étant purement technique.

Dans le courant de *l'information retrieval,* l'appariement est évalué mathématiquement. Le modèle booléen est historiquement le premier, le plus répandu malgré ces faibles performances. Les documents sont représentés par un ensemble de termes non pondérés ; la requête est représentée par une expression booléenne et l'appariement se résume à une « valuation de pertinence binaire » vraie ou fausse correspondant à la présence ou l'absence d'un terme, et la réponse fournie comporte un ensemble de documents (Gaussier et Stéfanini, 2003 : p. 33). Il existe des extensions de ce modèle (modèle booléen pondéré notamment). Le modèle vectoriel introduit par Gerard Salton dans les années 70 représente les mots significatifs des documents et des requêtes par des vecteurs dans un espace euclidien (Besançon, 2004 ; Memmi, 2000). Chaque élément du vecteur est un réel qui correspond au poids du mot donné dans le document et la requête, le calcul du poids étant fondé sur les fréquences d'occurrences. Suivant cette conception il n'y a plus d'ordre dans le texte qui est réduit à un « sac de mots ». La pertinence est évaluée en tenant compte des positions respectives d'un document par rapport à une requête et est estimée par une distance, c'est-à-dire au sens mathématique, une mesure définie dans un espace euclidien, qui traduit une proximité sémantique. Il existe plusieurs variantes de ce modèle dit « standard », certaines prenant en considération la notion de cooccurrence, c'est-à-dire des associations de termes. Des techniques de traitement de la langue plus ou moins évoluées sont appliquées aux documents pour sélectionner les termes d'indexation : élimination des mots-outils par l'application d'anti-dictionnaires, filtrage de l'information sur des critères linguistiques, réduction de l'espace de dimension par des méthodes mathématiques comme l'indexation sémantique latente (Latent Semantic Indexing).

Le modèle probabiliste repose sur le principe du classement probabiliste dont les caractéristiques ont été énoncées pour la première fois en 1977 (Robertson, 1977). Jian-Yun Nie et Jacques Savoy en présentent les principes généraux : *« Un document n'est pas simplement retourné ou non à l'usager mais on lui attribue une probabilité de pertinence (parfois nommée score ou degré de similarité comme dans le cas du modèle vectoriel) »* (Nie et Savoy, 2004 : p. 56). Dans cette approche, la pertinence est modélisée comme un événement probabiliste et donne à voir à l'utilisateur une liste ordonnée de documents. La recherche d'information est considérée comme un processus itératif mettant en jeu le principe de rétroaction dans lequel un utilisateur après avoir soumis une première requête, fournit au système les documents jugés pertinents ou non. La sélection des termes à inclure dans la représentation d'un document fait l'objet de différents modèles d'indexation probabilistes, le plus connu étant fondé sur la loi de Poisson qui décrit la distribution théorique des occurrences d'un mot dans un document, puis dans une collection. Il existe plusieurs mesures de calcul de la distribution des occurrences dans un ensemble de documents : loi de Pareto, loi de Zipf, modèle Bayesien, etc.

A la fin des années 90, émerge enfin une nouvelle famille de modèles fondée sur la logique et disposant d'un niveau de généralité théorique suffisant pour tenir compte des modèles de recherche d'information existants (van Rijsbergen, 1986) (Nie, 1989). L'objectif consiste toujours à modéliser la correspondance entre les documents et les requêtes, mais en considérant que chacune des représentations sur lesquelles opère le formalisme est dégradée. Ainsi, l'index ne représente-t-il que partiellement les documents et les requêtes ne sont qu'une approximation du besoin de l'utilisateur. Pour Jean-Pierre Chevallet, tenir compte de ces dégradations dans un modèle de recherche d'information permet d'expliquer la différence d'appréciation entre le jugement de pertinence du système et celui de l'utilisateur (Chevallet, 2004 : p108). Il existe plusieurs types de logiques, entendons par là des langages formalisés pour réaliser des déductions ou manipuler des valeurs de vérités (logique des propositions, logique des prédicats du premier ordre, logique multivaluée, logique modale, logique abductive, etc.), la question essentielle étant de prendre en considération l'incertitude (inhérente à la notion de « besoin ») et de la modéliser.

Quels que soient les modèles, les critères de mesure de performance sont le rappel et la précision[12] qui évaluent « objectivement » la pertinence d'un document vis-à-vis d'une requête : l'on parle de *pertinence système*[13]. Les protocoles d'évaluation ont été mis au point en laboratoire (*Laboratory-based Model*) (Cleverdon, 1967). Si le courant de *l'information retrieval* est dominé par les modèles mathématiques, ces derniers s'appliquent néanmoins à du matériau langagier. En effet, l'indexation automatique porte sur des textes, qui, traités par des ordinateurs, sont réduits à des chaînes de caractères. La transformation de ces chaînes de caractères en représentations codées et formalisées contribue à améliorer la pertinence système. **Par conséquent, dans ce cadre de recherche, la médiation technique impose ses contraintes, ses partis pris, ses objets de description sur la modélisation du langage**. Les traitements linguistiques de l'information ont pour objet de transformer les objets textuels fournis en entrée en des représentations exploitables par les machines. Malgré l'omniprésence de la technique, nos travaux ne relèvent que partiellement de l'ingénierie linguistique : d'une part, nos choix linguistiques ont contribué à faire évoluer l'environnement technique et n'ont pas seulement consisté à décrire et coder des données linguistiques. D'autre part, l'environnement technique était davantage considéré comme un cadre de recherche fédérateur pour les membres du laboratoire qu'un prototype de SRI. **En fin de compte, nos travaux sont le résultat d'un positionnement scientifique, à mi-chemin entre le développement d'applications pour la recherche d'information et les prises de position théorique sur le langage pour orienter des choix techniques.**

Cette première partie comporte deux chapitres. Le premier chapitre décrit les connaissances contenues dans les dictionnaires et les grammaires d'un analyseur morphologique dérivationnel destiné à reconnaître automatiquement les mots suffixés. Cet analyseur est l'un des modules du système de reconnaissance du français écrit (CRISTAL) destiné à indexer des textes intégraux. Le second chapitre montre que la prise en compte de connaissances linguistiques pour profiler les textes permet d'améliorer les méthodes de classification pour la recherche d'information.

[12] Le rappel (recall) mesure la capacité d'un SRI à retrouver tous les documents pertinents à une requête, la précision (precision) mesure sa capacité à ne retrouver que ces documents pertinents.

[13] System or algorithmic relevance (Saracevic, 1975)

1. Le langage comme modèle d'organisation des connaissances : *normaliser, interpréter et décompter*

Les notions clés de l'organisation des connaissances à l'œuvre dans ce chapitre peuvent se résumer ainsi : les connaissances que l'on utilise pour indexer les textes intégraux sont issues d'un modèle linguistique de la morphologie dérivationnelle du français. Ces connaissances ont pour vocation de normaliser et interpréter les formes graphiques issues des textes intégraux, ce qui permet leur décompte. Normaliser, interpréter et décompter sont les trois étapes préalables à la démarche d'indexation en texte intégral. Le nombre de publications qui portent sur ce thème est de 19 sur 45 qui se déroulent entre 1989 et 1999, mémoires de maîtrise, de DEA et de doctorat compris.

1.1. L'environnement de notre recherche : éléments de cadrage

1.1.1. Le TAL et la recherche d'information

Les modèles issus du traitement automatique du langage naturel (TALN) ou TAL[14], sont utilisés en complément d'autres types de modèles mathématiques (statistiques, probabilistes), améliorent, sous certaines conditions, les performances des systèmes de recherche d'information. Christian Fluhr, chercheur au C.E.A est à l'origine de l'un des SRI les plus renommés pour le rôle que jouent les technologies du langage : le système SPIRIT[15]. Ce chercheur indiquait dès la fin des années 80 que dans les SRI, les traitements linguistiques étaient quasi-inexistants sur l'anglais alors qu'ils étaient très poussés sur le français. D'après cet auteur, cette situation s'expliquait notamment en raison d'une plus grande complexité de la morphologie française par rapport à la morphologie anglaise. Il mentionnait que les traitements morphologiques

[14] En France, le consensus sur le terme de *Traitement Automatique des Langues* (TAL) date des années 90, l'expression Traitement Automatique du Langage Naturel (TALN) était également en vigueur. Aux Etats-Unis à la même époque, deux courants se côtoyaient qui finirent par se scinder : « the computational Linguistics » (traduit en français par *linguistique informatique*, parfois par *linguistique computationnelle*) est la discipline qui traite des aspects fondamentaux, alors que « the Natural Language Processing », se donne des objectifs d'applications industrielles en réponse à des demandes sociales (Cori et Léon, 2002 : p. 32-33)

[15] Christian Fluhr a développé SPIRIT, le premier SRI fondé sur une approche linguistique sophistiquée (Fluhr 1977, 1985, 1992).

contribuaient à déterminer les unités du langage dans les textes devant par la suite servir d'événements aux modèles statistiques dans les opérations de comparaison (Fluhr, 1985 : p. 99). D'après cet auteur, alors que le choix des modèles statistiques n'avait qu'une faible influence sur la qualité de l'interrogation, le choix des événements linguistiques sur lesquels allait s'exercer le modèle et donc, la qualité du traitement linguistique préalable, étaient en revanche, déterminants (*ibid*., p. 101).

Dès les années 60, un lien étroit se noue entre la linguistique et la recherche d'information en raison d'un objectif commun : la chasse aux ambiguïtés des langues naturelles. Le TAL se développe d'abord sous l'influence de la « Computational Linguistics » et appréhende la langue par niveaux (Fuchs, 1993) : la morphologie, la syntaxe, et la sémantique, dont les études se sont développées au fil du temps. Plus récemment, entre en compte le niveau discursif[16]. L'orientation du TAL est d'abord théorique et mobilise des modèles formels (Cori, 2008). Entre les années 80 et 90, c'est le niveau syntaxique qui concentre l'essentiel des travaux de recherche sous l'influence chomskyenne. La morphologie bénéficie de cette orientation, en particulier la morphologie flexionnelle qui entretient des liens étroits avec la syntaxe par le biais de ses marques de flexion (accord singulier / pluriel, marques de personnes, etc.) avec les fonctions syntaxiques. Le but des traitements linguistiques est d'identifier les unités discrètes des textes par segmentation, d'opérer des regroupements structurels puis d'attribuer des informations – ce que l'on nomme des « étiquettes », d'où la dénomination d' « étiquetage » – aux unités elles-mêmes ainsi qu'aux relations structurelles. Les enjeux de chaque niveau sont de réduire les ambiguïtés du langage naturel et d'attribuer du « sens » au moyen d'étiquettes, c'est-à-dire des symboles (Fuchs, 1993 ; Jacquemin et Zweigenbaum 2000).

En TAL, deux orientations sont considérées, l'analyse – ou reconnaissance – automatique et la génération automatique, que la nature des traitements permet de distinguer. Les définitions désormais classiques de ces orientations sont données dans les manuels d'introduction au TAL. Ainsi, Pierrette Bouillon propose-t-elle les définitions suivantes :

[16] (Péry-Woodley and Scott, 2006).

> *« L'analyse des langues naturelles vise à extraire des textes leur représentation. Elle prend donc comme entrée un texte et lui assigne une ou plusieurs représentations qui rendent explicites toutes les informations jugées pertinentes. […] [nous] définissons la génération comme l'opération inverse de l'analyse, qui part d'une représentation de texte et fournit en sortie un texte en langue naturelle. Ces deux types de transformations nécessitent diverses connaissances que l'on regroupe généralement en deux composantes distinctes, linguistique et informatique.*
>
> *(i) La composante linguistique comprend la description formelle de la langue, c'est-à-dire les connaissances linguistiques […]*
>
> *(ii) La composante informatique regroupe, quant à elle, les programmes informatiques ou algorithmes, qui spécifient comment interpréter la description de la langue. » (Bouillon, 1998 : p. 27-28)*

L'auteur souligne l'indépendance de la description linguistique et des algorithmes : c'est ainsi que la description des données linguistiques est abordée d'une manière déclarative indépendamment de l'ordre de traitements des programmes. La maintenance et la modification des données n'a ainsi pas d'impact sur la mise au point des programmes. Ces considérations achevées, évoquons les liens qui unissent la morphologie dérivationnelle et la recherche d'information.

1.1.2. La morphologie dérivationnelle et la recherche d'information

L'idée d'intégrer des traitements morphologiques propres à la dérivation est ancienne. Karen Sparck-Jones, chercheuse britannique reconnue pour ses travaux pionniers en traitement automatique des langues et recherche d'information, publie en 1974 dans la revue *Journal of Documentation* un article célèbre intitulé « Automatic Indexing » dans lequel elle présente les apports des racines de mots en particulier si leur description est fondée sur le plan sémantique et syntaxique (Sparck-Jones, 1974).

Dans la recherche d'information appréhendée comme un ensemble de techniques destinées à favoriser l'appariement, les traitements morphologiques peuvent porter sur les documents et / ou sur les requêtes. Par ailleurs, la morphologie peut être appréhendée sous l'angle de l'analyse ou de la génération automatiques. Dans le cas de l'analyse, il s'agit de reconnaître qu'une chaîne de caractères est un mot construit,

de le segmenter en ses constituants minimaux, de valider et d'interpréter le découpage ; dans le cas de la génération, il s'agit de « générer » tous les mots construits possibles (pas forcément attestés) à partir des constituants minimaux disponibles dans la langue. Par conséquent, l'appariement s'effectue de deux façons suivant que l'on appréhende l'index sous sa forme condensée, réduite à ses unités minimales ou suivant que l'on considère l'index sous une forme étendue, comportant l'ensemble des mots construits possibles de la langue. Le choix des index et le choix du statut des unités minimales engagent ici un point de vue sur la langue (cf. 1.2).

La reconnaissance automatique recourt à divers traitements tels que la segmentation, les prétraitements morphologiques, l'analyse morphologique, la désambiguïsation, l'analyse syntaxique, etc. L'analyse morphologique recouvre deux domaines de la linguistique, la morphologie flexionnelle et la morphologie dérivationnelle. Les « outils » qui traitent la morphologie « computationnelle » connaissent plusieurs appellations. Les outils les plus évolués sont les analyseurs morphologiques qui ont pour objectif de segmenter un mot en ses composants minimaux : flexions, suffixes, préfixes, etc. puis de valider et interpréter le découpage produit. Une seconde famille d'outils existe qui se contente de ramener une forme fléchie à sa forme canonique : les lemmatiseurs. Ces outils peuvent aussi ramener une forme dérivée à sa racine : les « *raciniseurs* », traduits de l'anglais « *stemmers* ». Dans le cas des raciniseurs, il n'y a pas d'interprétation du découpage réalisé, si bien qu'il s'ensuit de nombreuses ambiguïtés. Il s'agit plutôt de procédures permettant de regrouper les mots d'une même famille en mobilisant des techniques assez frustes comme « *des techniques de désuffixation et de recodage supprimant les affixes (essentiellement les suffixes) pour isoler les pseudo-racines* » (Moreau et Sébillot, 2005 : p. 6).

Dans l'optique de la génération automatique, l'appariement entre des requêtes et des documents consiste à *« élargir la description de la requête à tous les termes de la même famille afin de couvrir les différentes formes dérivées qui apparaissent dans le texte »* (Clavier, 1996a : p. 18-21). Ainsi, le principe d'extension de requêtes consiste à engendrer « *par reformulation* » les familles de mots liées à une même unité linguistique. Suivant l'option théorique retenue, l'unité de référence sera comme

indiquée *supra* les mots, les morphèmes, les racines, etc. afin d'élargir la recherche (Clavier, *ibid.*).

1.1.3. Le CRISS : un cadre de recherche pluridisciplinaire

Les travaux menés au CRISS sous la direction de Jacques Rouault s'appliquaient à l'informatique documentaire (Rouault, 1987). Ces travaux mettaient en œuvre le programme de recherche pluridisciplinaire du groupe SYDO[17] sur l'indexation automatique de documents textuels. C'est dans ce cadre qu'ont été menées à bien l'élaboration d'une grammaire morphologique flexionnelle du français (Berrendonner, 1983b) et la réalisation de l'analyseur morphosyntaxique pour la reconnaissance automatique du français écrit, dénommé CRISTAL (Antoniadis, 1984). Une réflexion déterminante sur le statut des unités d'indexation a été conduite par les participants du groupe SYDO, notamment par Michel Le Guern (1983) et par Jean-Paul Metzger dans sa thèse d'Etat portant sur les syntagmes nominaux et l'information textuelle (Metzger, 1988). Suivant leurs propositions, les unités d'indexation, les descripteurs, sont des unités du discours à caractère référentiel, de nature syntagmatique et non lexicale : ce sont essentiellement des syntagmes nominaux ainsi que certains adverbes, des marques flexionnelles verbales marquant le temps (Metzger, 1988 : p. 26-27). Ces entités syntagmatiques sont analysées dans une perspective logico-sémantique. Dans l'optique d'extraire automatiquement des syntagmes nominaux pour l'indexation, Geneviève Lallich-Boidin a développé un analyseur syntaxique du français puis testé le prototype de son algorithme sur un corpus de résumé d'articles scientifiques en botanique (Lallich-Boidin, 1986). Les travaux réalisés jusqu'à cette époque s'appliquent à l'information scientifique et technique, par exemple, la géologie (Veilex, 1985), et participent du vaste mouvement que d'aucuns ont appelé « l'informatisation de la société »[18] car ils oeuvraient à la

[17] Le groupe de recherche dénommé SYDO (pour SYstèmes DOcumentaires) réunissait le Centre de Recherches Linguistiques et Sémiologiques de l'Université Lyon II (Michel Le Guern), le Département de Linguistique Française de l'Université de Fribourg (Alain Berrendonner), Le laboratoire d'Informatique Documentaire de l'Université Lyon I (Richard Bouché), le Centre de Recherche en Informatique et Sciences Sociales de l'Université des Sciences Sociales (Jacques Rouault).

[18] Le terme d'informatisation a progressivement remplacé celui d'automatisation. Aujourd'hui, on parle de numérisation. Ces termes ne sont pas équivalents même s'ils sous-entendent des traitements informatiques.

mise en place de systèmes informatiques à l'hôpital, à l'université, dans des bibliothèques, des entreprises.

Au début des années 90, les préoccupations liées à l'informatique documentaire basculent vers celles de la recherche d'information en texte intégral. Le mémoire de DEA de Gilbert Eymard (1988)[19], puis sa thèse de doctorat (1992)[20] constituent un tournant : la question n'est plus de rechercher l'information à partir des représentations condensées des documents – résumés, notices bibliographiques – mais bien d'accéder au texte intégral en exploitant la structure logique des textes grâce aux sommaires et en s'appuyant sur une « organisation hiérarchico-logique des termes » pour rechercher l'information (Eymard, 1992 : p. 251). La thèse de Gilbert Eymard qui visait l'extraction de connaissances dans un manuel technique de l'une des versions du minitel, est emblématique des enjeux des politiques de l'information scientifique et technique du moment. D'une part, les demandes d'applications en termes de traitements automatiques des données textuelles conduisent à systématiser les partenariats entre les entreprises et les universités, d'autre part, le Centre National d'Etudes des Télécommunications (CNET), à travers son projet de télématique, apparaît comme le partenaire stratégique de l'industrialisation de l'information (Saläun, 1993).[21]

Dans les années qui ont suivi, les thématiques en lien avec l'informatique documentaire se sont effacées au profit d'enjeux de la recherche d'information en texte intégral. Le développement des bases de données textuelles est considéré

[19] Eymard Gilbert, *L'interrogation en langue naturelle d'une base de données textuelle : un chantier. Application aux « Spécifications Technique d'Utilisation du Minitel 1B (STUM1B) »*, Mémoire de DEA, CRISS, Université de Grenoble2, CENT/PPA/OGE, 1988.

[20] Eymard Gilbert, *Traitement documentaire des sommaires : des mots-clés à l'extraction de connaissances. Application à une documentation technique*. Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Grenoble2, 1992.

[21] On oublie trop souvent d'indiquer à quel point les universités ont joué un rôle important dans la recherche et la mise au point d'outils de traitement automatique. De nombreux laboratoires ont élaboré d'excellents prototypes, mais leur maintenance et leur mise à niveau dans des langages de programmation qui évoluent en permanence sont difficiles à assumer en l'absence de moyens. Deux solutions sont possibles : confier le prototype à une société privée qui dépouille alors l'université de ses inventions et la fait ensuite contribuer financièrement pour acquérir l'outil, ou le laisser mourir (ce qui a été le cas de l'analyseur morphologique développé dans l'équipe Cristal).

comme l'étape qui a précédé le développement des moteurs de recherche[22]. Le laboratoire s'est alors pleinement inscrit dans le paradigme technique « orienté système ». Cette posture signifie, ainsi que nous l'avons déjà indiqué, que les travaux qui relèvent de cette orientation résultent d'un découpage des objets de recherche calqués sur les composantes des systèmes de recherche d'information. Tous ces objets ont été traités par des chercheurs et des doctorants issus de disciplines différentes : informaticiens, psychologues, documentalistes, linguistes. L'ensemble des travaux était pensé à la manière d'un puzzle dans lequel chacun avait sa place[23], en voici quelques exemples :

- en lien avec le développement d'interfaces : la thèse de Christel Froissart sur la robustesse des interfaces[24] ; celle de Catherine Chanet sur la notion de « demande » dans un environnement de dialogue homme–machine[25] ;
- en lien avec le développement de formalismes de représentation des connaissances : la thèse de Mounia Fredj sur les modèles logiques de raisonnement et notamment les formalismes orientés objets fondés sur la logique lesniewskienne[26] ; ou la thèse de Jean-Marc Francony sur la prise en compte du contexte d'interaction pour représenter les connaissances dans un système de dialogue multi-modes[27] ;

[22] *« Search engines are structurally similar to database systems. Documents are stored in a repository, and an index is maintained. Queries are evaluated by processing the index to identify matches which are then returned to the user. However, there are also many differences.[…]»* (Zobel and Moffat, 2006 : p. 2)

[23] On parlait beaucoup de « chantiers » et quasiment toutes nos activités de recherche étaient en chantier en raison de l'interdépendance des projets autour d'un même objet.

[24] Froissart Christel, *Robustesse des interfaces homme-machine en langue naturelle*. Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1992.

[25] Chanet Catherine, *La demande dans le dialogue finalisé : de la surface linguistique aux représentations de l'action.* Thèse de doctorat en sciences de l'information et de la communication sous la direction de Jacques Rouault, Sciences de l'éducation, Université Grenoble3, 1996.

[26] Fredj Mounia*, Saphir. Un système d'objets inférentiels : contribution à l'étude des raisonnements en langue naturelle.* Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

[27] Francony Jean-Marc, *Modélisation du dialogue et représentation du contexte d'interaction dans une interface de dialogue multi-modes dont l'un des modes est dédié à la langue naturelle écrite*. Thèse de doctorat en informatique, sous la direction de Jacques Rouault Université Pierre Mendès-France, Grenoble2, 1993.

- en lien avec le choix et le développement d'architectures supportant le système de reconnaissance automatique du français CRISTAL : la thèse de Marie-Hélène Stefanini[28] pour les questions d'architecture multi-agents et celle de Karine Warren[29] sur la gestion des conflits entre les agents ;
- en lien avec la génération automatique du français, les thèses de Laurence Balicco[30] et de Claude Ponton[31] ;
- en lien avec l'activité de recherche d'information d'experts en situation de travail la thèse de Céline Paganelli[32], et la thèse d'Evelyne Mounier[33] sur la notion de paragraphe et la question de l'unité d'affichage à restituer à l'utilisateur en réponse à une requête ou à générer automatiquement.

D'autres travaux pourraient être cités, nous n'avons mentionné ici que ceux qui ont été contemporains des nôtres et avec lesquels il y avait le plus d'interaction. Après cette présentation du contexte scientifique, venons-en à nos propres travaux.

1.2. L'apport de la linguistique à l'indexation automatique

Le recours à la linguistique pour indexer les textes, entendons par là normaliser et documenter les textes avant de les soumettre à des calculs statistiques, était un axe fort du programme scientifique de Jacques Rouault qui affirmait en 1987 :

[28] Stefanini Marie-Hélène, *Talisman : une architecture multi-agents pour l'analyse du français écrit*. Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

[29] Warren Karine, *Gestion de conflits dans une architecture multi-agents d'analyse automatique de texte.* Thèse de doctorat en sciences de l'information et de la communication sous la direction de Jacques Rouault, Université Grenoble3, 1998.

[30] Balicco Laurence, *Génération de répliques en français dans une interface homme-machine en langue naturelle.* Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

[31] Ponton Claude, *Génération automatique de textes en langue naturelle : essai de définition d'un système noyau*. Thèse de doctorat en sciences de l'information, sous la direction de Jacques Rouault, Université Stendhal, Grenoble3, 1996.

[32] Paganelli Céline, *La recherche d'information dans des bases de documents techniques. Etude de l'activité des utilisateurs*. Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, à l'Université Stendhal, Grenoble3, 1997.

[33] Mounier Evelyne, *Etude expérimentale de la segmentation d'un texte en paragraphes*. Thèse de doctorat en sciences de l'information, sous la direction de Jacques Rouault et d'André Bisseret, Université Stendhal, Grenoble3, 1996.

> *« Notre hypothèse de travail est que l'analyse d'un énoncé en langage naturel ne peut s'opérer sans faire appel à des fondements théoriques. La linguistique est la science la plus apte à proposer ses modèles pour des données de nature textuelle [...] le recours à la linguistique est le seul guide sûr dans le passage des formes de surface au codage recherché : seules les procédures linguistiques introduisent dans la démarche une rigueur suffisante pour catégoriser, regrouper et interpréter. » (Rouault, 1987 : p. 68)*

Aujourd'hui encore, nous considérons que cette assertion est parfaitement fondée, en tout cas si l'on parle de rigueur de description et de modélisation, l'unique objection que l'on puisse émettre à notre sens, réside dans la mise en œuvre informatique : la recherche d'unités de rang inférieur au mot conduit à segmenter les mots, ce qui produit de nombreuses ambiguïtés liées à la multiplicité des découpages possibles. Or, la mise en œuvre de procédures de contrôle des opérations de segmentation nécessite souvent des informations de haut niveau.

Le travail que nous avons réalisé sur la morphologie dérivationnelle dans le cadre de la thèse de doctorat, du stage post-doctoral et des publications réalisées entre 1993 et 1999 comporte quatre aspects :

- une approche théorique du domaine de la morphologie dérivationnelle linguistiquement fondée conduisant à une proposition de modélisation en TAL ;
- une intégration du modèle de la suffixation dans la chaîne de reconnaissance du français écrit avec des perspectives d'implémentation pour améliorer l'indexation automatique ;
- une évaluation des résultats de l'analyse morphologique dérivationnelle sur des corpus proposés dans le cadre de la campagne d'évaluation Grâce ;
- deux applications destinées à valider le modèle : la première propose une étude en corpus pour extraire des connaissances dans des documents techniques ; la seconde est une application de la morphologie dérivationnelle à l'analyse automatique des discours.

Un informaticien « taliste » serait allé jusqu'à la formalisation et à l'implémentation d'un prototype[34]. La démarche que nous avons suivie dans notre thèse était de nature expérimentale et consistait à s'inscrire dans un modèle général de la recherche d'information – *a laboratory model of IR* suivant (Ingwersen and Järvelin, 2005 : p.4). Cette forme de contextualisation « en laboratoire » consiste à identifier et à définir précisément les contraintes du SRI qui ont une incidence sur un phénomène étudié, dans notre cas, la modélisation de la morphologie dérivationnelle. La section suivante est consacrée à l'élaboration du « cahier des charges » de l'analyseur morphologique d'un système de recherche d'information défini en laboratoire.

1.2.1. Nos choix d'indexation

1.2.1.1. Prépondérance du texte intégral

Aujourd'hui l'accès au texte intégral semble une évidence. Mais il n'en a pas toujours été ainsi. Dans les années 90, la recherche d'information en texte intégral était devenue une priorité pour la recherche appliquée, notamment le TAL, et un terrain d'investigation possible pour les chercheurs depuis que les bases de données textuelles étaient diffusées sur des supports magnétiques. Néanmoins, la disponibilité de ces documents était toute relative, ce qui explique que plusieurs travaux de doctorat conduits au CRISS ont recouru au même corpus pour valider leurs travaux, le fameux, et ô combien poétique manuel d'utilisation du logiciel informatique G-COS fourni par la société Bull.

Les enjeux de ce type de recherche sont présentés dans notre thèse comme fort différents de ceux de la recherche documentaire et de la recherche d'information factuelle dans des bases de données. La recherche documentaire avait fait l'objet de travaux antérieurs au CRISS et nous en décrivions les objectifs dans notre thèse : *« la sélection [d'un document] se fait à partir de la description signalétique des documents et d'une représentation condensée de leurs contenus : indexation sous la forme de mots-*

[34] Dans les années 90, l'enseignement du TAL n'était pas très répandu contrairement à aujourd'hui, où les étudiants bénéficient d'une double formation en linguistique et informatique dès la licence. Pour ma part, j'étais plus linguiste qu'informaticienne.

clés, de descripteurs, et, éventuellement, sous forme de résumés » (Clavier 1996a, p.6). La recherche d'information factuelle, dont les prolongements actuels résident dans les systèmes de question-réponse, faisait référence dans les années 90 *« à certains services télématiques qui renseignent sur les horaires de train, les cours boursiers, les numéros des téléphone »* (*ibid.,* p.7). Cette distinction avait pour but de poser les principes d'un nouveau type de recherche d'information qu'autorisent aujourd'hui les moteurs de recherche sur internet : écrire une requête dans une interface sous la forme de mots-clés, obtenir des documents en réponse et accéder au texte intégral. La seule différence réside dans le fait qu'au début des années 90, la recherche d'information portait sur des données textuelles issues d'univers fermés, des bases de données, alors qu'aujourd'hui, les données proviennent du web, un environnement ouvert.

Dans ce contexte de recherche d'information en texte intégral, nous avons décrit le rôle que joue la morphologie dérivationnelle pour rapprocher les requêtes des documents *via* les index.

En premier lieu, nous avions fait le choix de travailler en reconnaissance et de limiter le domaine de description à la suffixation. Nous voulions proposer un modèle qui serve à développer un analyseur : un modèle évolué (en termes de niveaux de description) qui segmente, valide et interprète le découpage en respectant des principes linguistiques. Nous nous sommes placée en reconnaissance et avions écarté la génération, considérant que cette application ne pouvait être envisagée qu'à partir du moment où la morphologie serait précisément décrite. Des références de publications récentes montrent que l'orientation en génération est beaucoup plus active qu'il y a quinze ans, ce que nous interprétons comme une marque de maturité de la morphologie computationnelle. Ces travaux ont pour objectif le développement de ressources lexicales pour le TAL. Ainsi, Aurélie Merlo (2012), doctorante en linguistique[35], propose-t-elle un système de prédiction de néologismes des noms en *–ier* afin de compléter automatiquement les lexiques pour le TAL. Cette approche est fondée sur une analyse linguistique fine des contraintes morphologiques. Mais il

[35] Thèse en cours, Merlot Aurélie, *De la nécessité des connaissances encyclopédiques en morphologie constructionnelle* sous la direction de Georgette Dal et Fiametta Namer.

existe d'autres approches, comme par exemple celle de Ludovic Tanguy et Nabil Hathout (2002) qui se servent du « web comme corpus » pour étendre les lexiques avec « des formes lexicales nouvelles » en fonction de leur terminaison. Les suffixes retenus portent sur des noms d'action en *-ade, -age, -ance, -ement, -ence, -erie, -tion*. Les auteurs prédisent, à partir du candidat « icônification », les formes « *icônifiait, icônifiant, icônifie, icônifient, icônifier*, etc. », ce qui conduit à surgénérer des séries dérivationnelles.

En second lieu, nous avions pour but de concevoir un modèle de la suffixation compatible avec le modèle linguistique de la morphologie flexionnelle définie par Alain Berrendonner en 1983 (Berrendonner, 1983b). Ce modèle présentait une double caractéristique, d'une part, il comportait peu de catégories morphosyntaxiques (10 classes majeures), ce qui en faisait un analyseur susceptible de s'adapter à différents types de discours (un analyseur généraliste). D'autre part, les catégories étaient définies pour leur pertinence syntaxique, si bien que les informations relatives aux variations formelles de la surface, étaient régularisées. Les classes majeures définies sur des principes distributionnels étaient sous-catégorisées afin de refléter d'autres types de contraintes (syntaxiques, lexicales).

Ces deux voies orientaient nos choix de formes canoniques représentant les séries dérivationnelles. Cependant, d'autres questions se posaient. Le représentant d'une famille dérivationnelle devait-il être un mot, un morphème, une racine, une lexie ? Quelles informations, i.e. quels traits linguistiques, devaient porter ces formes canoniques pour être compatibles avec la syntaxe et en même temps documenter le niveau morphologique ? Quelles étaient les règles qui décrivent la combinaison des unités internes à la morphologie ? Ces questions étaient posées dans notre thèse (Clavier, 1996a : p. 74). Nous précisions également que les traitements envisagés dans le cadre de la reconnaissance n'étaient pas de simples jeux de réécriture de la structure de surface puisqu'ils engageaient la signification des unités de langue et le sens lié aux opérations morphologiques de dérivation. En outre, la constitution d'index devait poursuivre des objectifs plus pragmatiques, tels qu'en réduire la taille,

tout en respectant des critères de spécificité[36] et d'exhaustivité, en normaliser les entrées tout en respectant les relations morphologiques propres au lexique d'une langue.

In fine, ces traitements linguistiques avaient pour objectif de rapprocher des contenus sémantiquement voisins lors de la recherche d'information. Ceci pouvait se concevoir au niveau des familles de mots, des emprunts à d'autres langues, de la synonymie, des variantes orthographiques, etc. ou au niveau syntaxique comme la recherche d'énoncés paraphrastiques. Ces enjeux étaient exprimés dans l'un de nos articles écrit en 1994 :

> *« Nous souhaiterions [donc] développer un modèle plus puissant, qui rassemble autour d'un lemme des familles présentant des variantes graphiques : par exemple, trucage et truquage doivent être identifiées par rapport au verbe TRUQUER ; la lemmatisation devrait également regrouper des formes allomorphiques telles producteur, production autour d'un verbe PRODUIRE. De surcroît, lorsqu'on rencontre des formes dérivées telles cécité, nous souhaiterions les considérer comme construites au même titre que acidité et avidité. Par conséquent, cette forme dérivée sera identifiée par un lemme, CEC-, bien que ce lemme soit contrairement aux précédents, privé d'autonomie syntaxique et qu'il présente une forme lemmatisée non "standard". Grâce à cette description, on pourra alors envisager dans un deuxième temps, de rapprocher les lemmes AVEUGLE et CEC par un lien de parasynonymie et ainsi notre modèle permettra de traiter les cas de supplétions qui sont de fait, laissés de côté par toutes les méthodes fondées sur la comparaison de chaînes de caractères. Ces rapprochements nécessitent une analyse morphologique, ce qui pose alors le problème de la segmentation des unités linguistiques, de leur description et de leur représentation dans un système d'analyse du français pour la recherche d'informations. » (Clavier et Lallich-Boidin, 1994)*

1.2.1.2. Sélection de textes longs et spécialisés

En raison de la forte demande sociale qui pesait sur l'accès informatisé à la documentation d'entreprise, plusieurs travaux au CRISS portaient sur des textes très spécialisés comme la documentation technique. Ces textes présentaient la particularité d'être longs ce qui leur valait l'appellation de « gros corpus », dont le

[36] « *La spécificité caractérise la capacité à garder une information précise et non généralisée* » (Chartron et al., 1989)

sens n'était pas le même que celui que lui attribue la linguistique de corpus (voir 2.1.2, le deuxième chapitre de cette partie). Pour la recherche d'information, la longueur des textes posait deux contraintes supplémentaires à l'indexation. D'une part, il fallait définir la taille de la partie de texte à afficher à l'utilisateur. C'est pourquoi Evelyne Mounier s'était intéressée au rôle du paragraphe et d'autres travaux ont porté sur la question de la segmentation du texte à partir de critères autres que statistiques[37]. D'autre part, il convenait de définir la granularité de l'indexation, notion qui posait les questions du degré de finesse de l'analyse automatique et celle de la nature des constituants à indexer. La définition de la granularité de l'indexation est liée à l'application.

Dans le cas de la documentation technique G-COS, 520 pages consacrées à l'administration d'un environnement informatique nécessitent une indexation plus fine que celle qui repose sur le simple repérage de mots-clés, même lemmatisés, afin de localiser des sous-parties du texte et ne pas donner le document entier en réponse à l'utilisateur. L'exemple du document technique G-COS concentre un ensemble de questionnements centraux pour l'indexation de textes scientifiques et techniques. La documentation technique est conçue pour des usages très spécialisés nécessitant une analyse fine du texte. Elle comporte un vocabulaire technique répétitif, présentant peu d'ambiguïtés de langue. Dans ce cas de figure, une indexation thématique par « sujets » est insuffisante. Par conséquent, la seule identification de syntagmes nominaux se révélait inadaptée ainsi que le montrait Céline Paganelli dans sa thèse de doctorat : les techniciens en situation de maintenance d'un logiciel cherchent à identifier des procédures, à localiser des commandes, *etc.* c'est-à-dire des éléments langagiers qui ne sont pas uniquement des thèmes ou des sujets. De nombreux exemples de requêtes émaillent notre mémoire de thèse et illustrent les types de tâches auxquelles sont confrontés les techniciens face à un manuel d'utilisation d'un logiciel, e.g. « *Comment programmer la vitesse de la prise péri-informatique ?* » (*ibid*., p. 9). Nous en déduisions *« [qu]'il [fallait] supposer que dans le cadre de cette application qui [relevait] de la résolution de problèmes, les descripteurs [devaient] permettre*

[37] Ouerfelli Tarek, *La segmentation des documents techniques composites dans une perspective d'indexation. Vers la définition d'un modèle dans une optique d'automatisation.* Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Stendhal, Grenoble3, 2001.

d'identifier également des tâches à accomplir, ainsi que les divers « actants » et « circonstants » de leur accomplissement. »

La contribution de la morphologie dérivationnelle à l'indexation de textes techniques a été développée dans le cadre de notre stage post-doctoral, consacré notamment à l'analyse des noms construits dans la documentation technique (Clavier, 1998)[38]. Dans le manuel d'utilisation GCOS, les procédures et les descriptions d'objets techniques apparaissent dans le texte sous la forme de procès verbaux (*programmer la vitesse*) et de noms déverbaux (*la programmation de la vitesse*), ce qui nécessitait de s'intéresser plus spécifiquement à ces unités langagières. Or, durant la période de rédaction de notre thèse (1992-1996), la majorité des logiciels d'indexation automatique commerciaux ne s'intéressaient qu'aux syntagmes nominaux, en raison de leur portée référentielle : *« A l'heure actuelle, une tendance domine qui tend à assimiler les candidats-descripteurs aux syntagmes nominaux. C'est le cas des systèmes CLARIT (Evans & al. 1991), SIMPR (Smeaton, 1991) pour l'anglais et des systèmes SPIRIT et ALETH de la société GSI-ERLI pour le français. »* (*ibid*., p. 10). Par ailleurs, la caractérisation de textes techniques n'était également pas très répandue bien que la demande sociale fût très importante pour le TAL. Ce constat semble se confirmer aujourd'hui, où l'information professionnelle ne jouit pas d'une reconnaissance importante au sein de la communauté scientifique (Clavier et Paganelli, 2013). Nous citerons une exception : Marie-Paule Péry Woodley qui, avec d'autres collaborateurs (Virbel, Pascual) s'est intéressée à l'organisation textuelle des textes procéduraux (Péry-Woodley, 2001).

1.2.1.3. Prise en compte de la structure des textes

Dans notre thèse, nous indiquions que *« le traitement d'information en full-text était une pratique largement répandue dans la plupart des bases de données [actuellement] commercialisée (cf. les dépêches de l'AFP, les bases de données juridiques comme Lexis) »* (*ibid*., p. 8). En revanche, elle se pratiquait surtout sur du texte plat. Or, Gilbert

[38] CLAVIER V. (1998), *Etude sémantique des noms dérivés de verbe : problèmes d'aspect »* rapport final du stage post-doctoral - AUPELF-UREF, Université de Fribourg.

Eymard avait montré combien la prise en compte du sommaire était importante pour accéder à la documentation technique (Eymard, 1992). Dans la continuité de ces travaux, nous avions émis l'hypothèse (avec d'autres chercheurs) qu'il était nécessaire de prendre en compte la structure logique des textes longs : *« La taille volumineuse des documents suppose que puisse être prise en considération leur structuration logique,* i.e *les différentes subdivisions hiérachico-logiques du document en sections, chapitres, paragraphes, notes, renvois, titres. Un poids plus ou moins important peut ainsi être accordé à un descripteur en fonction de la place qu'il occupe dans l'organisation logique du texte »* (*ibid*., p. 7).

La contribution de la morphologie dérivationnelle se révélait particulièrement intéressante pour l'indexation des titres. L'étude du manuel d'utilisation GCOS indiquait que les titres du sommaire comportaient de nombreux mots construits, et en particulier des noms dérivés de verbes ou noms « déverbaux ». L'intérêt de la morphologie était alors d'extraire des connaissances en lien avec la recherche de procédures ou d'objets techniques (cf. la thèse de Céline Paganelli), puisque les suffixes que nous avions choisi de décrire permettaient de construire des noms d'action abstraits décrivant des procédures ou des noms d'objets concrets. Cependant, une grande partie des noms déverbaux en *–age, -ment –tion* qui apparaissaient dans le manuel présentaient une ambiguïté sémantique de type concret / abstrait parfois très difficile à lever hors contexte. De ce fait, un mot construit comme *programmation*, peut désigner une action, donc une étape d'une procédure (*la programmation de la vitesse*) ou un objet (*le bouton de programmation*).

1.2.1.4. Les connaissances dérivées des textes

L'indexation en texte intégral devait-elle s'appuyer sur des connaissances extérieures au texte, à partir de thesaurus ou d'autres langages (terminologies, vocabulaires, classifications etc.) ou sur le texte lui-même ? Telles étaient à l'époque de la rédaction de notre thèse, les questions qui se posaient à l'indexation automatique. Nous reprenions dans notre mémoire la distinction que proposait Hans Paijmans en 1993

sur l'indexation dérivée du document lui-même et l'indexation assignée[39] issue de connaissances extérieures. Hans Paijmans s'appuyait sur deux catégories de technologies, le logiciel commercialisé TOPIC qui recourait à une indexation assignée et le logiciel CLARIT qui recourait à une indexation dérivée.

> *« There are essentially two approaches to the creation and maintenance of this document or knowledge representation. One is to create a knowledge system in advance and assign the documents to it afterward : assigning indexing. The other is to derive the terms of the index language from the documents themselves : deriving indexing »*[40] *(Paijmans, 1993 : 384)*

Dans notre thèse, nous présentions les avantages et les inconvénients de chacune de ces approches. Nous avons argumenté pour le recours à des méthodes d'indexation dérivées du texte lui-même.

Nous avions fait le choix de méthodes d'indexation dérivée et nous avancions comme argument l'idée que cette forme d'indexation est plus fidèle aux textes et restitue « *la richesse lexicale des textes* », ce qui permet de prendre en considération « *les néologismes si fréquents dans les textes techniques* » (Clavier, 1996a : p. 11). Pour justifier nos choix, nous évoquions la souplesse et l'économie que cela représente lorsqu'il n'y a pas de thesaurus ou d'autres systèmes d'organisation de connaissances à maintenir (*ibid.,* p. 12). Cependant, la question de la néologie lexicale est complexe. Elle pose tout d'abord la question du rapport à la norme d'usage : la présence ou non d'un mot dans les dictionnaires de langue étant l'argument pour rejeter ou pour accepter cet usage. Cette position normative avait déjà été critiquée en 1936 par le linguiste Nyrop, dans le 3ème tome de sa *Grammaire Historique* à propos de la formation des mots :

---

[39] L'indexation par assignation est définie par Jacques Chaumier comme suit : *« Dans ces méthodes [par assignation], l'indexation est réalisée non plus par des termes extraits du texte, mais par des mots-clés issus d'un thesaurus préétabli. On retrouve ici la séquence de l'indexation manuelle : extraction de concepts, - traduction en termes d'un langage documentaire »*. (Chaumier, 1982 : p. 62)

[40] Ce que nous avions traduit par « *Il y a essentiellement deux démarches de création et de maintenance de la représentation des documents ou des connaissances ; L'une est de créer un système de connaissances a priori et de lui assigner les documents après coup : l'indexation assignée. L'autre est de dériver les termes du langage d'indexation à partir des documents eux-mêmes : l'indexation dérivée ».* (*ibid.* p, 10)

> *« Les mots nouveaux attirent ordinairement la critique et commencent souvent par exciter l'hilarité ou l'indignation ; sous ce rapport, ils partagent le sort de tout ce qui est nouveau. […] On est parfois tenté de se demander à quoi sert la fureur toujours renaissante des grammairiens contre les mots nouveaux. Les résultats de leurs agissements sont ordinairement minces ou plus que minces. Et à quoi servent toutes les prescriptions qui tendent à restreindre le nombre de néologismes, les autorisent dans certains cas et les condamnant dans d'autres ? Ce ne sont que des efforts inutiles, des coups qui ne portent pas – et ne peuvent pas porter […] Selon nous, les néologismes sont les résultats nécessaires et les marques infaillibles de la vitalité forte et saine de la langue, ou pour mieux dire, ils témoignent d'une imagination poétique et plastique toujours en éveil, d'efforts continuels pour rendre l'expression plus variée, plus nuancée, plus pittoresque. Il ne faut pas endiguer le flot des néologismes. (Nyrop, 1936 : p. 11)*[41]

Ensuite, la néologie pose la question du statut du corpus dans lequel on rencontre le néologisme : un néologisme pourra être considéré comme une création littéraire, un effet de style s'il est écrit sous la plume d'un écrivain célèbre ou… comme une aberration dans une copie d'élève. L'usage du web comme corpus est à ce titre très problématique ainsi que le souligne le linguiste Pierre Lerat à propos du recueil de la terminologie[42] puisqu'il ne permet pas de contrôler ses sources.

Enfin, la néologie lexicale pose la question du rapport à la langue. Dans un article commun rédigé avec Alain Berrendonner à la suite de notre stage post-doctoral, nous avons montré que le lexique est constitué d'ensembles plus ou moins productifs et que certains suffixes seulement peuvent prétendre augmenter le stock lexical d'une langue. Nous nous sommes intéressés au suffixe homonymique en –*age.* Dans cette série dérivationnelle, plusieurs sous-systèmes se côtoient, dont certains, comme les noms dénominaux en –*age* (*kilométrage, cordage)* se sont éteints alors que les noms déverbaux (*coiffage, collage*) sont très productifs. Nous avons montré qu'il y avait un lien entre ces deux systèmes, les suffixes déverbaux pouvant en partie remplir les mêmes fonctions sémantiques que les suffixes dénominaux. L'étude se termine par un

[41] cité par (Coret, 1994 : p. 19).

[42] *« Il ne saurait être question de traiter le Web comme un corpus, car c'est par nature un non-corpus, voire un anti-corpus. Il ne présente en effet aucune des caractéristiques attendues classiquement d'un corpus : il n'est ni homogène, ni clos, ni même stable. Il importe d'évaluer constamment les sources, faute de quoi on ne saurait faire du travail correct »* (Lerat, 2005 : p.5)

ensemble de propositions expliquant pourquoi « *une série peut abonder en mots impossibles, plutôt qu'en mots possibles* » (Berrendonner et Clavier, 1997 : p. 44) :

> *« Au total, la série des dérivés en -age2 a bien quelque chose de paradoxal. D'une part, on y décèle sans peine un pattern dérivationnel régulier, dont les produits présentent une compositionnalité sémantique relativement transparente. On s'attendrait donc à ce que cette forte systématicité favorise chez les locuteurs la perception de la série en tant que sous-système combinatoire « calculable », et lui assure une certaine capacité générative. Or, les néologismes en N-age2 brillent par leur absence. On est donc en présence d'un sous-système lexical à la fois régulier et improductif. L'existence de tels modules oriente vers une conception sédimentaire ou stratifiée du lexique, comme formé de diverses couches de formes construites différant notamment par leur degré de ritualisation ou de « figement » combinatoire. Les dérivés en N-age2 appartiennent visiblement à un stade intermédiaire entre les résidus compositionnellement opaques et les séries dérivationnelles majeures. » (Clavier et Berrendonner, 1997 : p. 43-44)*

La néologie pour le TAL est un problème épineux puisqu'il est l'une des causes d'échec majeure des analyseurs C'est ce que montrent Christel Froissart et Geneviève Lallich-Boidin lors de l'évaluation de l'analyseur CRISTAL (nous y revendrons en 1.2.3.1). La typologie des mots non reconnus par l'analyseur morphologique flexionnel révèle que la dérivation est « l'une des causes de nombreux rejets plus spécialement dans *Le Monde* » (Froissart et Lallich-Boidin, 1996 : p. 91). Dans les langues spécialisées qui se caractérisent par une terminologie foisonnante, le problème devient alors central. La couverture lexicale que peut offrir un outil de reconnaissance automatique est alors l'un des arguments déterminants pour faire face à la créativité du langage. On ne raisonne cependant plus uniquement en termes de néologismes mais en termes de présence / absence dans les dictionnaires quel que soit le statut du mot inconnu. Pour atteindre cet objectif de reconnaissance des mots inconnus, deux choix sont possibles : soit l'analyseur comporte un dictionnaire qui tend vers l'exhaustivité (approche qui était proposée par le Laboratoire d'Automatique Documentaire et Linguistique dirigé par Maurice Gross[43]) soit l'analyseur comporte un dictionnaire minimal et a la capacité de reconnaître les mots nouveaux. Nous avions adopté cette dernière position ce qui nous avait conduit à

[43] (Gross, 1975 ; 1994)

faire des propositions d'organisation des connaissances dans des dictionnaires de morphèmes (cf. 1.2.2.3).

Au-delà de ces prises de position pour le TAL, nous nous sommes toujours intéressée à la partie du lexique productive responsable de l'accroissement du vocabulaire d'une langue plutôt qu'aux hapax ou aux séries improductives. Ce point de vue a été défendu dans différents environnements et à plusieurs reprises dans notre activité de recherche et d'enseignement.

- Dans le contexte de la traduction automatique français-allemand, nous avions montré dans notre mémoire de maîtrise en 1989 quels sont les ressorts de régularité qui favorisent la productivité de séries adjectivales suffixées en *-bar*[44] (qui correspondent aux adjectifs en *–able* en français tels *mangeable, buvable*, etc.) Nous montrions que les procédés morphologiques français et allemands bien que tout aussi productifs l'un que l'autre ne s'appliquent pas forcément aux mêmes bases verbales, ce qui pose des problèmes de traduction.
- Dans notre mémoire de DEA, nous avions montré comment et pourquoi il pouvait exister des séries morphologiques très productives mais concurrentes. Ainsi, les suffixes en *–age, -ment* et *–tion*[45] *(ex. abattage / abattement)* permettent tous les trois de construire des noms dérivés de verbes, ce qui paraît peu économique au regard du système d'une langue. Mais dans les faits, l'examen de séries concurrentes met en évidence des spécialisations sémantiques (Clavier et Coret, 1998), ce qui présente un intérêt pour la génération automatique.
- Dans le cadre de notre monitorat, Roger Sauter, professeur de linguistique allemande à Saint-Etienne nous avait confié un cours de « lexicologie et morphosyntaxe » dans le module d'enseignement de linguistique destiné aux

[44] Clavier Viviane, *Untersuchung von Ableitungsregeln der bar-Adjektive im Hinblick auf maschinelle Übersetzung,* Mémoire de maîtrise d'allemand sous la direction conjointe de Roger Sauter et Geneviève Lallich-Boidin, Université Jean Monnet, Saint-Etienne, 1989.

[45] Clavier Viviane *Morphologie dérivationnelle des substantifs déverbaux,* Mémoire de DEA en sciences du langage, sous la direction de G. Lallich-Boidin, Université Stendhal, Grenoble 3, 1990.

étudiants de LEA et LLCE, uniquement consacré aux suffixes, préfixes et éléments de composition allemands. Ce cours était nouveau et permettait notamment de systématiser l'apprentissage du vocabulaire.

- Plus tard, dans le cadre d'un programme de recherche pluri-formation, nous nous sommes intéressée à l'acquisition de compétences lexicales dans le cadre scolaire de l'enseignement en primaire[46] ; nous sommes également intervenue dans un cours de didactique pour l'enseignement du vocabulaire à l'Université d'Orléans devant un public d'étudiants de sciences du langage, mention didactique du français langue maternelle[47].

1.2.2. Nos choix théoriques de modélisation pour le TAL

Au moment où nous rédigions notre thèse, il régnait en France une forte dynamique autour de la morphologie dérivationnelle grâce au SILEX, laboratoire lillois dirigé par Danielle Corbin, qui, quelques années auparavant avait rédigé une thèse imposante intitulée « *Morphologie dérivationnelle et structuration du lexique* », en deux volumes publiée aux éditions Max Niemeyer Verlag (Corbin, 1987). Ce travail avait été pour nous une révélation, tant par le domaine de description considéré – un ouvrage entièrement consacré au lexique – que par l'ampleur de la bibliographie traitée et les méthodes d'analyse employées. Nous avions rencontré plusieurs fois Danielle Corbin qui nous avait mise en contact avec une autre doctorante, Muriel Coret résidant à Paris, qui faisait une thèse de doctorat sur les mêmes suffixes que nous, sous la direction d'Hélène Huot à Paris 7[48]. A partir de 1993, nous avons assisté plusieurs années de suite aux séminaires d'Hélène Huot dans lesquels plusieurs doctorants travaillant en morphologie se retrouvaient régulièrement. Nous avons ainsi été introduite dans le cercle des morphologues que nous avons rencontrés à l'occasion de séminaires à Paris 7, de journées d'études (de l'ATALA ou de l'AFLA), de colloques

[46] Programme Pluri-Formation « Produire des textes en situation d'apprentissage : rôle des compétences lexicales, textuelles et cognitivo-pragmatiques ». dir. C. Golder 2004-2007.

[47] Ce cours de 2h n'est pas mentionné dans mon curriculum viate (intitulé « Lexique et vocabulaire : quelques enjeux théoriques et didactiques », Licence SDL, mention DFLM, Université d'Orléans.

[48] Coret Muriel, *Problèmes de suffixation et structuration du lexique. Etude des mots en –eur, -age, -ment, -ion.* Thèse de doctorat en linguistique sous la direction d'Hélène Huot, Université Denis Diderot, Paris 7, 1994.

(*Les forums de morphologie*), d'écoles d'été ou de congrès (par exemple *International Congress of Linguists).*

Dans la théorie linguistique[49], la morphologie bénéficie d'une longue histoire qui remonte à l'Antiquité (Bourquin, 1975) mais ce n'est véritablement qu'au XIXème siècle avec la grammaire historique que sont apparus les premiers travaux d'envergure notamment le *Traité de la formation de la langue française* du comparatiste Arsène Darmester en 1890. Cependant, ces travaux s'inscrivent dans une perspective historique qui met l'accent sur l'origine et l'évolution des mots plutôt que sur les propriétés des processus morphologiques reconnus en synchronie, perspective adoptée depuis Saussure. Négligée par la grammaire générative qui, entre les années 60 et 50 privilégie la syntaxe, le « retour » de la morphologie remonte aux années 70 avec les travaux de Morris Halle (1973), Peter Matthews (1974) et Mark Aronoff (1976). Ces trois auteurs sont connus pour leurs travaux en lien avec la grammaire générative, Morris Halle, en phonologie générative, les deux autres en morphologie. Il ressort de ces travaux que le lexique est plus régulier qu'il n'y paraît[50], ce qui confère à la morphologie un statut autonome dans le modèle génératif. La perspective générativiste a donné lieu d'abord à un traitement transformationnel qui a connu un grand succès en France notamment avec les travaux de Dubois et Guilbert qui ont inspiré une série de dictionnaires édités chez Larousse[51]. Cette perspective appréhende la suffixation en termes de transformations de phrases de base. Elle a été vivement critiquée parce qu'elle ne reconnaît pas la spécificité du lexique. Par exemple, ne sont pas prises en compte les lacunes lexicales (*le *fumage de la cigarette est interdit*), les cas de lexicalisation qui font que le sens prédictible n'est pas le sens attesté (*-ifier* ne signifie pas toujours « rendre Adj ». Ex. *justifier*) (Coret, 1994 : p. 32). Les théories transformationnelles ne reconnaissent pas non plus les mots dérivés apparaissant dans des contextes différents (on ne dit pas une *tranche hépatique* pour dire que l'on mange une *tranche de foie*). C'est ainsi que progressivement, la morphologie s'est imposée comme un domaine à part entière,

[49] Nous nous appuyons sur l'introduction de la thèse de Muriel Coret qui dresse un panorama de la place de la morphologie dans la linguistique (p. 15-39).

[50] Dans les modèles de la grammaire générative, le lexique a un statut idiosyncratique.

[51] *Dictionnaire du français contemporain* (1971), *Grand Larousse de la langue française* (1971), *Lexis* (1975).

mais elle a dû se faire une place sur un double terrain : celui de la phonologie et de la syntaxe. Ces questions se sont posées à nous lors de l'intégration du module de traitement de la morphologie dans la chaîne du français :

- pour l'interaction phonologie / morphologie, cette question a été traitée lors de l'établissement des règles d'alternance formelle des suffixes (ex. *professeur / professorat*) et résulte d'un travail de recension systématique des alternances phono-graphiques de notre corpus (Clavier et Lallich-Boidin, 1993).
- pour l'interaction syntaxe / morphologie, cette question a été discutée au moment de se prononcer sur le statut des morphèmes. Nous avons fait le choix d'un traitement unifié de la syntaxe et la morphologie, considérant le principe d'économie et de généralité de la langue : ainsi les unités de rang inférieur au mot sont catégorisées comme noms, adjectifs, verbes, *etc*. ce qui permet de décrire une syntaxe interne au mot (Clavier et Lallich-Boidin, 1994).

Enfin, les sections suivantes répondent à la question de savoir comment définir un mot construit, comment recueillir les unités de rang inférieur au mot, comment les nommer et comment décrire les règles de formation.

1.2.2.1. Définition d'un mot construit

Nous avons adopté la perspective de Danielle Corbin pour définir les mots construits. (Corbin, 1991). *« Nous adoptons une perspective morpho-sémantique de la dérivation qui consiste à retenir comme construit un lexème donc les constituants sont formellement et sémantiquement interprétables »* (Clavier, 1996a : p. 62). Cette position s'inscrit dans le courant que Danielle Corbin (1987, 1991) qualifie de morphologie associative et prend ses racines dans les travaux de (Dell 1970 ; Halle, 1973 ; Aronoff 1976). Cette morphologie s'écarte des principes d'une morphologie non compositionnelle, comme celle développée par Claude Gruaz (1988) qui fonde

l'identification des segments sur des principes distributionnels excluant (dans un premier temps) le sens[52].

Il y a plusieurs registres terminologiques pour décrire les mots construits. En linguistique, l'unité qui nous intéresse est le morphème, c'est-à-dire le plus petit élément de signification. Mais tous les morphèmes n'ont pas le même statut, et pour certains linguistes le terme de morphème est réservé aux éléments grammaticaux. D'où le recours à la dénomination de lexèmes, pour les distinguer des morphèmes (grammaticaux). Les méthodes distributionnelles et le TAL ont en commun de mettre en évidence par les tests de commutation ou par troncation, des segments de mots, qu'on ne sait plus comment nommer : ainsi, les « troncats » ou « segments » lorsqu'on veut rester neutre, les « bases » qui peuvent être autonomes, courtes ou longues, les « racines » qui sont issues du latin ou du grec, les « joncteurs », qui sont les éléments de jonction entre une base et un suffixe *(anim-**at**-ion)*, *etc*. Dans la thèse, nous avons utilisé le terme de « base » conformément à la terminologie usitée dans le modèle morphologique flexionnel du système CRISTAL : une « base » est la partie fixe d'une chaîne de caractères qui est l'entrée d'un dictionnaire. Dans le modèle morphologique, une « base » n'a plus de statut dans le dictionnaire, elle désigne « le segment qui reste lorsqu'on a privé la chaîne de caractères de son suffixe » (Clavier, 1996a : p. 148). La base peut donc être simple ou construite, s'il reste encore des affixes, courte ou longue, s'il y a un joncteur, non autonome ou autonome, si elle connaît une réalisation sous forme de « mot ».

1.2.2.2. Corpus, méthode, résultats

L'ensemble des travaux réalisés sur la morphologie met en œuvre une approche descriptive du lexique construit par dérivation. La méthode est distributionnelle. Nous avons procédé à une analyse systématique de plusieurs séries de mots sélectionnés dans un dictionnaire électronique à partir de leurs terminaisons qui ne correspondent pas forcément à des suffixes (cf. par exemple *rade* versus *limonade*

[52] Nous avons également travaillé avec Claude Gruaz en 1998 à la révision de son dictionnaire synchronique des familles de mots (DISFA).

versus *carbonnade*). Nous nous sommes appuyée sur le DELAS[53] qui offre une couverture lexicale proche du *Petit Robert*.

> *« Pour réaliser l'étude linguistique, nous avons travaillé sur les données d'origine non encore interprétées, à savoir le dictionnaire DELAS du LADL (Gross, 1989) qui était à l'origine du dictionnaire Cristal. Le DELAS fournit 50.000 entrées lexicales sous leurs formes lemmatisées et la présentation très compacte de ces données nous a permis de sélectionner facilement les formes lemmatisées nominales et adjectivales terminées par les chaînes suivantes : -age, -aire, -eur, -ier, -ment et –tion. » (Clavier, 1996a : p. 123)*

Nous avons ainsi analysé 8.434 mots qui se répartissent comme suit : 1443 mots terminés par -*age* ; 613 mots terminés en –*aire* ; 1848 mots terminés en –*eur* ; 1155 mots terminés en -*ier* , 1247 mots terminés en –*ment* ; 2128 mots terminés en –*tion*. Les résultats de l'analyse sont en annexe de la thèse et constituent un travail de description linguistique significatif qui a pris plusieurs mois. Cette analyse met en évidence différents cas de figure conformes à la typologie proposée par Danielle Corbin et qui révèle un continuum entre des lexèmes construits et des lexèmes simples.

La typologie tient compte à la fois des variantes formelles qui peuvent affecter la base (*étrangler / strangulation*) ou le suffixe (*tuteur / tutorat)* ainsi que des cas de lexicalisation qui peuvent affecter la base alors qu'elle est pourtant formellement identifiable (cf. *professer* et *instituer* ont des sens éloignés de *professeur* et *instituteur*) ; ou des faits de spécialisations sémantiques des suffixes. La recension des cas d'idiosyncrasies suffixales met en évidence des sous-systèmes devenus improductifs : par exemple *dentition, foliation, nervation* ont une valeur collective idiosyncratique par rapport à l'ensemble de noms en –*tion* ; les noms féminin en –*eur clamer / clameur, errer / erreur* qui décrivent un état sont une exception parmi les suffixes en -*eur* construits sur des verbes, etc. Toutes les alternances formelles ont été répertoriées et ont fait l'objet de typologies justifiées sur le plan phonographique (*percer / perçage*), morpho-phonologique (*fleur / floral*) et morphologique (les différents statuts des segments joncteurs : *lion-c-eau, spasm-od-ique, annonc-iat-eur,* etc.) (Clavier et Lallich-Boidin, 1993).

[53] Dictionnaire électronique des mots simples du LADL.

Ce travail descriptif a permis de définir le statut et la forme des unités morphologiques ainsi que leur interprétation. Trois niveaux d'interprétation et de codification sont associés (Clavier et Lallich-Boidin, 1994) :

- le niveau morphologique permet de regrouper les formes graphiques de surface autour de représentants morphologiques. Les critères de regroupement sont phonologiques et l'on tient compte du statut des bases pour la flexion. A ce niveau, il peut y avoir une seule forme sous-jacente susceptible de décrire une série d'allomorphes : par exemple un seul suffixe –IER pour *cordonn-ier, bouch-er, boulang-er* ou trois bases possibles pour un lexème : PRODUIS- et PRODUI- PRODUC-, la dernière ne servant qu'à la dérivation (*producteur*).
- le niveau syntaxique permet à la fois de ramener les différentes variations d'un morphème ou d'un lexème à une unité syntaxiquement autonome. Ce qui présente l'avantage d'expliquer les cas de remotivation de racines comme *reptation / repter* versus *ramper*. Et, par ailleurs les suffixes sont appréhendés comme des opérateurs de catégorisation. Ainsi, les noms en *–eur* construisent des noms à partir de verbe : *chanter / chanteur*, mais aussi des noms d'état à partir d'adjectifs : *blanc / blancheur*.
- Le niveau sémantique regroupe les éléments du lexique qui présentent un lien sémantique au sein de séries dérivationnelles sans que le lien formel soit justifiable : ce sont ici les cas de supplétion lexicale, qui permette de traiter les cas de para-synonymie comme *cécité / aveuglement* ou les cas d'homonymie.

1.2.2.3. Modélisation pour le TAL

Il existe plusieurs façons d'aborder la morphologie pour le TAL. Les méthodes les plus connues ne sont malheureusement pas les plus évoluées linguistiquement. Ainsi, les algorithmes de Porter (1980) et Lovins (1968) sont-ils connus pour leur « stemmer »

ou « raciniseur »[54], qui, à partir d'une liste de suffixes donnés, voire de terminaisons, consistent à amputer les mots… et à regarder ce qui reste. Geneviève Lallich-Boidin a fait la démonstration du carnage que cela provoque sur le français, notamment avec l'algorithme de Lovins : *frilosité / fril*, *évidemment /évid*, *indépendamment / indépenda*… (Lallich-Boidin et Maret, 2005 : p. 41). Ces méthodes, sans connaissances linguistiques produisent des erreurs, y compris pour une langue comme l'anglais dont la morphologie est simple, alors qu'il est reconnu que la prise en compte des variantes morphologiques permet d'améliorer le rappel et la précision (Moreau et Sébillot, 2005 : p. 11). D'autres modèles ont été développés pour les langues à morphologie riche comme « la morphologie à deux niveaux », développée par Kimmo Koskenniemi (1983) sur le finnois. Ce modèle permet de mettre en correspondance une forme de surface avec une forme lexicale au moyen de règles. La morphologie à deux niveaux a connu un vif succès dans plusieurs langues dont le français (Namer et Schmidt, 1997). Lauri Karttunen que nous avions rencontré à Grenoble (Xerox) avait été l'un des premiers à implémenter le système Kimmo dans les années 80 pour l'anglais, le japonais, le roumain et le français (Beesley and Karttunen, 2003). Nous avions opposé des critiques à ce modèle à la fois sur le plan informatique et linguistique (Clavier et al, 1996).

En 1994, nous indiquions que contrairement à la morphologie à deux niveaux qui privilégiait « l'*indépendance de la modélisation par rapport aux langues* » et « *la bidirectionnalité* » (Clavier et Lallich-Boidin, 1994 : p. 132), nous faisions le choix d'une approche ancrée dans la langue française et visant uniquement l'analyse. La reconnaissance automatique devait répondre à deux objectifs. D'une part, il s'agissait de développer un modèle morphologique compatible avec le système CRISTAL et d'autre part, le modèle devait respecter trois types de contraintes pour « *orienter la segmentation vers une interprétation plausible* » (*ibid.* p. 132), i.e ainsi que nous l'avions proposé :

[54] *« Stemming has been the most widely applied morpholgical technique. With stemming, the searcher does not need to worry about the correct truncation point of searc keys. Stemming also reduces the total number of distinct index entries. Further, stemming causes query expansion by bringiing variants, derivations included, together. »* (Ingwersen and Järvelin, 2005 : p. 154).

- des contraintes de forme liées au statut morpho-phonologique des constituants internes du mot ;
- des contraintes syntaxiques liées à la catégorie grammaticale des constituants internes du mot ;
- des contraintes sémantiques liées à l'interprétation des bases et de l'opération affixale.

Dix ans plus tard, en 2000, Georgette Dal et Fiametta Namer déploraient qu'il n'existât toujours pas d'outils, analyseurs et générateurs, proposant une analyse sémantique des mots construits qui permettent l'expression des requêtes sous forme de termes et de relations de prédication entre les termes (Dal et Namer, 2000 : p. 429). Ces auteurs ont fait des propositions dans ce sens.

1.2.3. Organisation des connaissances dans un environnement distribué

La proposition de modèle de la suffixation que nous avions faite, constituait *« une première étape pour développer une application en TAL »* (Clavier, 1996a : p. 245). Trois aspects avaient été pris en considération : la place de la morphologie dérivationnelle dans l'analyse ; la description du module de traitement morphologique ; l'analyse des problèmes que soulève l'intégration du module morphologique au sein d'une architecture informatique.

Dans la chaîne de traitement du français écrit du système CRISTAL, les modules de reconnaissance coïncidaient avec les niveaux d'analyse linguistique ce qui conduisait à une approche segmentée des phénomènes langagiers, une approche dite « par niveaux » (Fuchs, 1993). En ce qui concernait la morphologie, il fallait envisager un traitement de la morphologie flexionnelle et de la suffixation. Si bien que le but de l'analyse morphologique était d'analyser une chaîne graphique, de la décomposer en ses constituants que ce soit des suffixes et/ou des flexions, et d'interpréter le découpage. Cette analyse hors contexte produisait de nombreuses solutions, pour lesquelles il n'était pas toujours possible de lever les ambiguïtés.

La question des ambiguïtés est le problème central du TAL : plus une analyse est fine, plus elle produit d'ambiguïtés. En voici quelques exemples :

- ambiguïtés liées au statut de la terminaison : entre un suffixe (limonade) ou une simple finale *(rade) ;* entre un mot qu'il faut segmenter (*avocaillon*) et un autre qui n'est pas construit (*papillon*) ;
- ambiguïtés entre plusieurs découpages possibles : *rations* est-il une forme conjuguée du verbe *rater* ou le pluriel du nom *ration* ?
- ambiguïtés de construction des mots : *garage* doit-il être interprété comme *coiffage*, i.e un nom d'action ? Et que dire de *plumage* est-ce un nom dérivé d'un verbe *plumer* ou d'un nom *plume* comme *kilométrage* dérivé de *kilomètre* ? (Berrendonner et Clavier, 1997)

Ces quelques exemples illustrent la complexité des phénomènes en jeu en se limitant à la description d'un seul domaine de la morphologie, la dérivation par suffixation. En effet, plus on opère de traitements sur le langage, plus on engendre d'ambiguïtés, et plus il faut de connaissances pour les lever. Il semble que ce soit une limite majeure du TAL, ainsi que le note Christian Jacquemin dans l'introduction de la revue TAL consacrée au traitement automatique des langues pour la recherche d'information :

> *« [En résumé], la convergence du TAL et de la RI n'est pas simple, elle suppose de faire les bons choix technologiques du côté du TAL pour avoir un réel impact sur les performances de la RI – les mauvais choix peuvent induire des performances plus faibles avec un coût calculatoire élevé. Il convient d'éviter les méthodes mettant en oeuvre des représentations des connaissances riches dont l'adaptation en grandeur nature est incertaine. » (Jacquemin, 2000 : p. 328)*

L'une des solutions auxquelles travaillaient les chercheurs informaticiens du CRISS était de redéfinir l'environnement informatique, afin de briser l'ordre séquentiel des traitements (Stéfanini, 1993) et (Warren, 1998). Il s'agissait de faire « coopérer » les modules ou de « gérer » leurs conflits en cas d'ambiguïtés au sein d'une architecture informatique distribuée. Au lieu de « modules », il était question « d'agents », dotés d'intention et d'autonomie qui « interagissaient » et « communiquaient » dans un « environnement » informatique. Ces travaux s'inspiraient de l'intelligence artificielle distribuée. Une publication résulte de ce travail en commun (Clavier et *al.* 1996). Cette collaboration scientifique ponctuelle nous a permis de préciser le contenu de l'agent morphologique et d'envisager les interactions avec l'agent morphologique flexionnel et la syntaxe. De cette façon l'intégration de l'analyseur morphologique

dans la chaîne de traitement CRISTAL était abordée en lien avec l'architecture du système.

Ainsi, dans sa partie interne, l'agent morphologique comportait-il une partie statique et une partie dynamique. La partie statique « *recouvr[ait] toutes les connaissances et compétences sur le savoir-faire linguistique de l'agent* » et la partie dynamique *« produi[sait] les activités de raisonnement et d'action, comme les stratégies représentant les différentes heuristiques linguistiques.* » (Clavier, 1996 : p. 246-260). Les connaissances morphologiques se situaient dans la partie statique de l'agent et comportaient :

- Les dictionnaires communs à la flexion et la dérivation (dictionnaires de lexèmes, compatibilités des suffixes et des flexions) ;
- Les dictionnaires spécifiques à la suffixation ;
- Les dictionnaires spécifiques à la flexion ;

En outre, les connaissances mettaient en œuvre des traitements qui pouvaient être décrits au moyen d'une grammaire régulière d'ordre 3, dans la classification chomskyenne. La description précise des connaissances morphologiques a fait l'objet d'une publication commune avec Muriel Coret (Clavier et Coret, 1997). Cet article se focalise sur la structure et le contenu de dictionnaires électroniques de morphologie destinés à être associés à des applications qui relèvent du traitement automatique des langues : traduction automatique, recherche d'information. Cette collaboration a permis une confrontation sur un certain nombre de points théoriques et a permis de mettre en commun les données, soit environ 8000 mots analysés sur le plan morphologique.

La réflexion sur les dictionnaires constitue un point nodal du TAL et de l'organisation des connaissances. Sur ce point, le Laboratoire d'Automatique Documentaire et Linguistique (LADL) du CNRS créé par Maurice Gross en 1972 à Jussieu (Université de Paris 7)[55] bénéficiait d'une expérience importante, le LADL ayant développé une série

[55] http://infolingu.univ-mlv.fr/LADL/Historique.html

de dictionnaires électroniques (DELAS, DELAF, etc.) pour le TAL[56]. Avec son équipe, Maurice Gross avait lancé un vaste programme de description systématique des propriétés syntaxiques du lexique ayant conduit à la confection de « lexiques-grammaires ». Ces derniers étaient à l'origine de nombreux projets de dictionnaires électroniques et de bases de données lexicales (Courtois et Silberztein 1989). S'appuyant sur les théories distributionnelles et transformationnelles de Zellig Harris, les lexiques-grammaires recensaient en grandeur réelle les structures syntaxiques et morphologiques élémentaires. Toutes les formes devaient être représentées dans des tables de propriétés normalisées et codifiées, et ne comportaient aucune information culturelle, étymologique et sémantique. L'ensemble des données et des programmes utilisés pour la reconnaissance ou la génération automatique était représenté par des automates d'états finis[57]. Suivant ces principes, la notion de néologisme n'avait aucune raison d'être, puisqu'une occurrence lexicale n'avait pas d'existence en-dehors d'une phrase support[58]. Par ailleurs, la présence d'une phrase support dans le lexique-grammaire était soumise à des critères d'acceptabilité syntaxique de locuteurs « compétents » (au sens chomskyen du terme) et non à des critères de jugements esthétiques ou d'écart à la norme d'usage. Concernant les connaissances morphologiques, Blandine Courtois signalait que la perspective appliquée s'inscrivait dans une optique de recension « tendant vers l'exhaustivité» des unités lexicales et de « codification systématique et rigoureuse » entre autres, des propriétés morphologies des mots (Courtois, 1994-1995).

Nous nous inscrivions en faux par rapport à cette position ainsi que nous l'indiquions dans nos travaux :

> *« Le fait de consigner toutes les formes lexicales (simples et complexes) conduisait à un lexique hétérogène qui présentait le double inconvénient d'être redondant — il ne distinguait pas le simple du construit – et non hiérarchisé – il mettait sur le même plan le régulier et l'irrégulier. Cette position comportait enfin l'inconvénient de réduire le lexique à une liste, nécessairement close à un moment donné, ce qui ne rendait pas compte de la créativité lexicale ». (Clavier et Coret, 1997 : p. 308-309)*

[56] Le DELAS avait d'ailleurs été donné à notre laboratoire et avait permis de réaliser l'analyseur Cristal.

[57] http://infolingu.univ-mlv.fr/LADL/Historique.html

[58] Pour Chomsky, le lexique contraint les transformations syntaxiques.

Dans cet article, nous proposions au contraire de :

> *[…] « définir un lexique non redondant, constitué d'unités élémentaires. Ce lexique sera l'entrée de règles morphologiques dont le rôle est de produire le lexique possible de la langue, non fini par définition. Ce sont ces règles qui prennent en charge la hiérarchisation des unités, en traitant l'opposition entre le régulier (produit par la règle, non consigné) et l'irrégulier (non prédictible, consigné). » (ibid.)*

Nous reprenions alors :

> *« la distinction entre lexique restreint et lexique étendu telle qu'elle est établie dans (Fradin, 1993b : 14-15) à la suite de (Zwicky, 1992) qui oppose un lexique défini en intension, "le noyau du lexique dénué de toute redondance pouvant servir de base à l'inventaire complet de lexèmes de la langue si on se donne un ensemble approprié de règles morphologiques" à un lexique défini en extension, qui comporte "l'inventaire total, mais possiblement infini, des lexèmes". » (ibid.)*

Ces prises de position orientaient vers le choix d'un modèle calculatoire.

1.2.3.1. Evaluation du modèle de la dérivation suffixale

Ainsi qu'il est d'usage dans le domaine de l'*information retrieval*, les systèmes de recherche d'information sont soumis à une évaluation qui permet de mesurer le taux de rappel et de précision. Parmi les campagnes d'évaluation, TREC est la plus renommée internationalement et celle qui a connu la plus grande longévité, bien qu'elle ne soit pas la première[59] : aujourd'hui, elle traite un grand nombre de langues et de thématiques et bénéficie de soutiens institutionnels puissants, malgré les critiques qui lui sont opposées. TREC, pour Text REtrieval Conference[60] i.e la conférence sur la recherche d'information textuelle, a été initiée en 1992 aux Etats-Unis à l'initiative du NIST (National Institute of Standards and Technology) et de l'agence DARPA (Defense Advanced Research Project Agency). Cependant, à ces débuts, c'est surtout l'anglais qui était privilégié.

[59] Salton indique que les premières évaluations remontent aux années 50 mais que c'est à partir des années 60 qu'elles ont été financées par la National Science Foundation (Salton, 1991).

[60] http://trec.nist.gov/

La campagne GRACE (Grammaire et Ressources pour les Analyseurs de Corpus et leur Évaluation) en revanche portait sur l'évaluation des analyseurs morpho-syntaxiques du français. Le projet était à l'initiative du LIMSI et de l'INaLF et avait débuté en 1994 dans le cadre du programme SHS-SPI du CNRS « Cognition, Communication intelligente et Ingénierie des langues' ». Il s'était achevé en 1998 sous l'égide du programme « Ingénierie des Langues (IL) »[61] Ce programme s'était déroulé en trois étapes (cf. Les archives du LIMSI):

1) *l'entraînement* pendant lequel un corpus brut d'environ 10 millions de formes en provenance à parts égales de la base FRANTEXT de l' INaLF et du journal *Le Monde* a été distribué aux 21 participants pour calibrer leurs systèmes ;
2) *les essais*, pendant lesquels 17 participants ont testé le protocole complet d'évaluation en marquant un corpus d'environ 450 000 formes ;
3) *les tests* au cours desquels 13 participants ont marqué un corpus d'environ 650 000 formes.

Le CRISS avait participé à la campagne d'évaluation de la robustesse de l'analyseur morphologique flexionnel CRISTAL à partir des corpus proposés aux participants (Froissart et Lallich-Boidin, 1996). Les résultats de cette analyse, et en particulier, la liste des mots qui n'avaient pas été reconnus avaient été analysés sur le plan linguistique, ce qui avait permis de valider notre modèle (Clavier, 1996b).

Nous montrions dans cette étude que 32% des mots inconnus étaient des mots construits par affixation (suffixation et/ou préfixation), par conversion et composition. Le corpus auquel avaient été appliqués ces tests comportait des textes littéraires du XIXème siècle issus de la base de données Frantext :

- GYP, *Souvenirs d'une petite fille*, Tome 1, 1927, 25100 mots.
- GYP, *Marie-Antoinette de Mirabeau, comtesse de Martel dite GYP*, 1849-1932, 48700 mots.
- Alexandre DUMAS, *La dame aux Camélias*, 1848, 69900 mots.

[61] Ces informations proviennent du site d'archives du LIMSI : http://archives.limsi.fr/RS99FF/CHM99FF/TLP99FF/tlp10/ (consulté en avril 2013)

- Charles NODIER, *La fée aux miettes*, 1831, 65100 mots.
- George SAND, *Histoire de ma vie*, 1855, Tome 1, 57300 mots.
- Alphonse de LAMARTINE, *Voyage en Orient,* Tome 1, 1835, 63000 mots

et une seconde série de textes rassemblait des articles du journal *Le Monde* (110.428 mots).

Après avoir identifié les modes de construction des mots construits, nous avions dressé un bilan des procédés en jeu, le parenthésage figurant les opérations propres à la syntaxe du mot (Clavier, 1996b : p. 104).

préfixation, suffixation :	*enrégimentement : [[en- [régiment]$_V$] -ment]$_N$* *impliable : [im- [[pli]$_V$ -able]]$_{Adj}$*
suffixation et composition :	*[[neuro]$_N$ - [[psyche]$_N$-ique]$_{Adj}$]$_{Adj}$*
suffixation, conversion et préfixation :	*inopéable : [in- [[O.P.A]$_{Vop}$-] -able]$_{Adj}$*
composition, préfixation, suffixation :	*crypto-antigaulliste :* *[[crypto]$_N$ - [anti- [[de Gaulle]$_{Npr}$ -iste]]$_{Adj}$]$_{Adj}$*

Cette expérience mettait en évidence la nécessité d'intégrer la dérivation dans les systèmes de reconnaissance automatique en raison du pourcentage important de mots construits inconnus. Ces mots inconnus étaient principalement issus de la presse, un corpus de textes spécialisés aurait fourni un pourcentage plus élevé. Cette étude pointait également la nécessité de prendre en considération l'ensemble des faits de dérivation, et pas seulement la suffixation. Or, ainsi que nous l'indiquions, ceci augmentait significativement la complexité des phénomènes à décrire, puisqu'il fallait envisager de recourir à des grammaires contextuelles (d'ordre 2). Les problèmes soulevés concernaient également le traitement des ambiguïtés.

Malgré l'intérêt qu'avait présenté cette confrontation aux données réelles, cette campagne d'évaluation posait néanmoins un problème de taille : celui du choix des corpus pour évaluer la pertinence des résultats d'un système de reconnaissance du français. Mobiliser des corpus du XIXème siècle en littérature pour évaluer des outils destinés à traiter en synchronie la langue générale, n'était guère recevable. L'existence de corpus n'était pourtant pas un fait nouveau dans les années 95, l'INaLF

ayant constitué un fonds de littérature important dès les années 60. Cependant, ces corpus étaient peu accessibles. Le fait que l'INaLF accepte de mettre ses données à notre disposition avait été, à l'époque, très apprécié par la communauté de chercheurs, et par notre laboratoire en particulier. Cependant, l'évaluation, telle qu'elle avait été menée dans le projet GRACE, était révélatrice de l'absence de réflexion qui présidait à la constitution des corpus en TAL. Benoit Habert et ses co-auteurs en avaient fait très justement le constat dès la fin des années 90 :

> *« Se trouvent parfois agrégés des documents avant tout parce qu'ils sont faciles d'accès : leur mise en relation n'a pas été réellement pensée. C'est ce que l'on pourrait appeler des regroupements « opportunistes ». Ainsi la communauté du traitement automatique des langues appelle souvent corpus les grandes collections de documents qui lui servent à mettre au point ses traitements. On serait tenté de voir là « du texte », texte dont on ne sait pas toujours très bien de quels usages langagiers il est représentatif. Pour nous, un corpus résulte d'un regroupement raisonné, conduit par une hypothèse de recherche explicite » (Habert et al. 1998 : p. 35)*

C'est pourquoi, il n'était pas étonnant, que lors de l'évaluation des résultats de l'analyse flexionnelle, Christel Froissart et Geneviève Lallich-Boidin eussent constaté que les mots inconnus révélassent « plusieurs états de langue » (p. 90). En effet, se trouvaient rassemblés des phénomènes qui ressortissaient à des faits d'oralité – dans les pièces de théâtre –, des variantes graphiques issues du vieux français, des citations en langues étrangères, des abréviations, des formules mathématiques, des noms propres, etc. L'ensemble de ces phénomènes conduisait à notre avis à remettre sérieusement en question la notion de « langue standard » comme niveau de représentation adéquat pour organiser les connaissances : à quoi pouvait renvoyer cette dénomination, si ce n'est à une conception du français contemporain écrit… vue au filtre des dictionnaires de langue ?

Ces critiques rejoignent d'autres limites théoriques et méthodologiques émises plus récemment sur les protocoles d'évaluation des performances des systèmes définis en laboratoire. Par exemple, Stéphane Chaudiron et Madjid Ihadjadene (Chaudiron 2004a, 2004b ; Ihadjadene et Chaudiron 2008) mentionnent les limites qui portent sur le protocole d'évaluation lui-même, notamment la construction des référentiels et le bien-fondé des métriques pour rendre compte de la satisfaction de l'usager :

l'absence de l'usager dans le processus d'évaluation, la modélisation simplifiée à l'extrême des pratiques informationnelles, la simplification de la notion de pertinence. Dans le cas de l'évaluation menée par la campagne GRACE, outre l'absence de prise en compte de l'usager, l'absence de réflexion sur la constitution de corpus était également à déplorer.

Malgré ces limites, la découverte de l'empirisme et des corpus nous étaient apparus particulièrement enthousiasmants, alors que le « TAL théorique » (Cori, 2008) issu du courant chomskyen, en perte de vitesse, était plus préoccupé de développer des modèles formels (GPSG, HPSG , etc.) que de se confronter aux données langagières attestées.

1.2.3.2. Morphologie dérivationnelle et extraction de connaissances

Lors d'un stage post-doctoral[62] réalisé en Suisse à l'Université de Fribourg sous la direction d'Alain Berrendonner, nous avons cherché à ancrer socialement le dispositif technique et avons proposé d'inscrire notre réflexion dans le cadre d'une application : l'extraction de connaissances dans des documents techniques. Deux directions ont été suivies. La première direction consistait à prendre le document technique comme objet d'étude tout en tenant compte de son utilisation en contexte professionnel et de sa dimension langagière. Le document lui-même était qualifié de corpus, voire même de « gros corpus ». La seconde orientation présentait une incursion en sémantique, l'extraction de connaissances nécessitant de rechercher des informations précises qui ne pouvaient être exprimées sous la forme de mots-clés, ou de chaînes de caractères dans une requête.

La première direction s'était dessinée à la suite de la découverte des corpus comme données attestées servant à mettre à l'épreuve le modèle morphologique. Cette confrontation aux données langagières permettait de comprendre que la réalisation d'outils à large couverture était une utopie, puisque la « langue standard » comme système idéal et idéel, déconnecté de son essence sociale ne permettait ni de définir les contours lexicaux d'une langue, ni d'expliquer la variation et la créativité

[62] Nous avons bénéficié d'une bourse d'excellence de l'Aupelf-Uref en 1997.

langagière. Par ailleurs, l'introduction de méthodes fondées sur les corpus (*corpus-based approach*)[63] avait sonné le glas des exemples forgés issus de l'intuition des linguistes. Dans le même temps, d'autres voix influentes dans la communauté du TAL et de la recherche d'information exprimaient des doutes sur la contribution du TAL aux tâches traditionnelles de la recherche d'information (Sparck-Jones, 1999)[64]. Ces auteurs indiquaient que les seules applications qui pouvaient tirer bénéfice du TAL étaient l'extraction de connaissances (ou extraction d'information), les systèmes de question-réponse, le résumé automatique et la catégorisation de textes (Condamines et Poibeau, 2008).

Ces constats, qui n'étaient pas formulés aussi clairement à l'époque, nous avaient conduites à définir précisément les objectifs de l'extraction de connaissances dans le manuel d'utilisation du logiciel GCOS. Ce travail a été mené en collaboration avec Céline Paganelli et Christel Froissart et a donné lieu à une publication collective (Clavier et al., 1997). Nous indiquions alors quelles étaient les caractéristiques de ce corpus, le domaine concerné et les utilisateurs potentiels :

> *« The corpus we work on is the instruction manual of GCOS7[65] (operating system). This large document, written in French, is readable on CD-ROM which represents 861 Mo. Besides, this document is highly structured as it organized in 5 hierarchical level (categories, volumes, chapters, sections and paragraphs. The technical domain concerned is computer science, thus, the expert users are computer scientists. » (Clavier et al. 1997)*

Ces utilisateurs, experts en informatique, formulaient certains types de demandes liés à des tâches de maintenance ou d'installation du logiciel GCOS. Exprimées sous la forme de requêtes, ces demandes avaient été recueillies lors d'une expérimentation réalisée par Céline Paganelli auprès des experts eux-mêmes, soit 40 informaticiens interrogés avec des méthodes issues de la psychologie cognitive (Paganelli, 1997). La typologie des requêtes présentait deux types de demandes : des demandes d'informations de type procédural (*comment faire pour…* ?) et de type descriptif ou définitoire (*qu'est-ce que … ? à quoi sert … ? de quoi se compose*… ?) (Paganelli, 2012, p.

[63] Qui s'oppose à la méthode guidée par les corpus (*corpus-driven approach*).

[64] « *It is not clear, [either,] that NLP is required for some tasks that are closely related to ordinary retrieval* ». (Sparck-Jones, 1999) cité par Jacquemin (2000 : p. 327).

[65]We specially thank Bull SA, Paris which has put this corpus at our disposal and enables us to work on.

79-80). Par conséquent, la question qui se posait était de connaître les catégories linguistiques correspondant à ces demandes qualifiées de demandes de type ACTION ou OBJET, et de savoir où elles apparaissaient dans le document. La perspective devenait onomasiologique jusqu'alors, elle avait été sémasiologique. Par ailleurs, elle prenait en considération une dimension de la textualité : la structure logique. Pour limiter le domaine de description en langue, nous nous étions concentrée sur les faits de morphologie, bien que nous ayons identifié d'autres marqueurs langagiers susceptibles d'être extraits, dans les titres et dans le corps du texte.

Il ressortait de l'étude des correspondances entre les catégories cognitives et les marqueurs de surface : les ACTIONS étaient actualisées par les infinitifs, les formes jussives et différents marqueurs indiquant la séquentialité.

> *« as far as ACTION is concerned : presence of infinitive and jussive forms, mention of the user through the thematic-role as Agent, processes interpretable as accomplishment, and several marks indicating sequentiality. » (Clavier et al., 1997)*

Quant aux OBJETS, ils apparaissaient en position thématique et se combinaient avec des verbes d'état.

> *« as far as OBJECT is concerned : marks appear often at the topic position and OBJECTS are described by verbs which denote states. » (ibid.)*

Les mots construits par dérivation apparaissaient en nombre important dans le document : *l'utilisateur, le chargement, l'environnement*, etc. Cependant, lorsque les noms étaient suffixés en *–age, -ment* et *–tion*, deux interprétations étaient possibles : soit comme un OBJET (*l'environnement d'exploitation*) soit comme une ACTION (*le chargement de données*). La nominalisation s'était alors révélée problématique dans le cas de cette étude, puisque même dans un corpus spécialisé censé réduire les ambiguïtés, elle manifestait une polysémie entre les catégories OBJET et ACTION. Ainsi, étions-nous à nouveau confrontée à la résistance du langage naturel face à la modélisation.

La seconde direction que nous avons empruntée est liée à la possibilité d'investiguer plus avant le niveau sémantique en lien avec la morphologie dérivationnelle. Trois

publications (Clavier, 1997 ; Berrendonner et Clavier, 1997 ; Clavier et Coret, 1998), une communication (Clavier, 1999) et un rapport de recherche (Clavier, 1998) rendent compte de cette incursion en sémantique que nous avons tenté de conduire à la fois sur les plans fondamental et appliqué. Le recours à un niveau de description sémantique était lié à des exigences plus importantes en termes de caractérisation du contenu pour l'extraction de connaissances. Ainsi que nous l'avons vu précédemment, extraire des connaissances de type OBJET ou ACTION nécessitait de rechercher des parties de textes indexées suivant des critères qui ne sont ni des « mots » ni des « morphèmes » à proprement parler, mais des descriptions de tâches cognitives qu'il fallait qualifier sur le plan linguistique. Ce fut l'objet de notre travail post-doctoral en Suisse.

Le séminaire qui nous a accueillie pour ce stage avait à son actif de nombreuses études en grammaire et sémantique formelles d'une part (Berrendonner, 1983a) et sur le syntagme nominal et la nominalisation d'autre part (Apothéloz et Reichler-Béguelin, 1995 ; Berrendonner et Reichler-Béguelin, 1995). Pour Alain Berrendonner qui doutait que *« les mots aient réellement un sens* » et « *n'osait plus parier sur les chances de la sémantique véric*onditionnelle » (Berredonner, 1987 : p. 287), la représentation sémantique des « descripteurs nominaux » devait d'un côté faire le lien entre le signifié d'un prédicat complexe dont on s'attachait à définir la structure combinatoire interne et, de l'autre, proposer des éléments de modélisation formelle (Berrendonner, 1995 : p. 9). Pour cet auteur, le contenu sémantique d'un syntagme nominal était divisé en deux parties (*ibid.* p. 9) :

- des composantes instructionnelles qui indiquaient certaines opérations à exécuter sur les référents cognitifs partagés par les interlocuteurs. Ces instructions étaient marquées par des prédéterminants (un/le/ce).
- des composantes descriptives dont les signifiants n'étaient pas des unités atomiques mais bien des substantifs et leurs diverses expansions.

Ce cadre de description était éloigné d'une conception du sens où « les mots sont des atomes désignant des objets du monde réel » répertoriés dans des classes référentielles ; où l'on présuppose que « l'interprétation d'un mot se réduit à

l'identification du concept auquel il se rapporte »[66], et où le traitement sémantique est envisagé pour « lever l'ambiguïté des mots de la requête et du document en leur associant un ensemble de sens non ambigus » (Valette et Slodzian, 2008 : p. 121). Cette conception était pourtant à l'oeuvre dans les ressources développées pour le TAL dans les années 90, comme WordNet par exemple (Voorhees, 1993), qui décrit les rapports conceptuels entre les mots (rapports de synonymie, antonymie, méronymie…). Aujourd'hui, les ontologies ont pris le relai de cette conception référentielle du monde.

1.2.3.3. Incursion en sémantique : études de cas

L'application envisagée pour extraire des connaissances à partir de manuels d'utilisation avait mis en évidence une difficulté importante liée à la sémantique des noms construits. Les noms dérivés de verbes présentaient une polysémie de type abstrait / concret qui limitait fortement la contribution de la morphologie à l'extraction de connaissances. En outre, certaines séries morphologiques présentaient également des cas d'ambiguïtés retorses aux traitements automatiques : l'homonymie suffixale. Homonymie et synonymie suffixale sont deux cas d'ambiguïtés typiques mettant la langue à l'épreuve de la machine : l'homonymie provoque du bruit et la synonymie, du silence.

Deux études ont été conduites durant notre stage post-doctoral sur ce sujet. La première étude portait sur la description des procédés de nominalisation dits « concurrents ». Elle a donné lieu à une publication (Clavier et Coret, 1998). La seconde portait sur l'analyse d'une série suffixale homonymique, les noms en *–age* dérivés de noms et de verbes, et a été présentée au premier Forum de morphologie organisé à Lille en 1997 (Berrendonner et Clavier 1997). Dans les deux cas, il

[66] Cette position est à l'origine d'une conception du lexique comme nomenclature et dont les prolongements se manifestent dans la création d'ontologies. *« On en reste généralement à l'idée, bien antérieure à la formation de la linguistique, que les langues sont des nomenclatures (et non des corpus de textes produits dans des pratiques différenciées) : leur lexique serait une représentation de choses qui, comme l'affirmait déjà Aristote, sont « les mêmes pour tout le monde ».. Cette conception antique informait la scolastique, et, par le biais d'auteurs comme Stuart Mill, elle a été reformulée dans la philosophie du langage qui sert aujourd'hui de cadre conceptuel au cognitivisme orthodoxe. »* (Rastier, 2004b)

s'agissait d'apporter quelques éléments de réponse sur les connaissances sémantiques propres à décrire des procédés morphologiques.

La première étude concerne les noms en *–age, -ment* dérivés de verbes *(chantage, ralliement)* et les noms dérivés sans suffixe *(sauter / saut).* Ces mots construits relèvent de procédés concurrents parce qu'il y a *« plusieurs procédés morphologiques possibles pour une même opération dérivationnelle »* (Clavier et Coret, 1998) : tous les noms dérivés sont des noms d'action qui peuvent avoir un sens concret – désigner un objet du monde – ou abstrait – décrire un processus, un événement, une action. Cette étude invitait à questionner le processus de nominalisation dans la même perspective que le passif, c'est-à-dire « *une opération réductrice de valence verbale* » (Berrendonner, 1995 cité dans Clavier et Coret, 1998), au sens où un verbe, lorsqu'il est nominalisé perd ses actants, exprimés sous la forme d'arguments. Cependant, certains d'entre eux peuvent être réactivés en position d'expansion nominale : il s'agissait de savoir lesquels. Il nous fallait un corpus pour sélectionner un corpus de groupes nominaux. Nous avons recouru au corpus de la campagne d'évaluation GRACE, à savoir les 6 romans du XIXème siècle, les articles du *Monde* des années 1989-1990. Nous avions rajouté le manuel GCOS et Muriel Coret avait également un corpus extrait du *Robert Electronique.* Il était difficile de parler de corpus, puisqu'il n'y avait aucun critère qui présidait à leur constitution, la seule qualité de ces documents était qu'ils étaient disponibles sur support électronique.

Notre hypothèse était que ces procédés n'étaient pas équivalents, l'existence de nombreux doublets – 460 pour les noms en *–age* et *–ment* comme *abattage, abattement* – prouvant que ce n'étaient pas des synonymes. L'étude conduisait à typer sémantiquement les expansions de ces noms soit comme des arguments du verbe d'origine, soit comme un complément de détermination du nom : par exemple, l'*accord du professeur* versus l'*accord de musique*. Lorsque l'expansion était ramenée à un argument du verbe, alors l'argument pouvait être interprété en termes d'actants, ou de « cas profond » ainsi que l'avait proposé Fillmore[67]. Lors de la dérivation, certaines opérations suffixales résorbaient irrémédiablement des actants. Ainsi la

[67] Nous avions en DEA travaillé à la réalisation d'un dictionnaire syntaxique des verbes (Clavier, 1990) décrivant les principales constructions verbales avec leurs arguments (suivant la méthode de Claire Blanche-Benvéniste).

suffixation en –*ment* résorbait l'Agent. Par exemple, dans *l'abattement de Paul*, *Paul* subit plutôt qu'il n'agit ; inversement, la suffixation en –*age* activait l'Agent. Dans *l'abattage de l'arbre*, le syntagme est perçu comme un procès agentif et l'on peut aisément introduire l'Agent : *l'abattage de l'arbre par Paul.*

Venait ensuite la question de l'application : l'étude de la nominalisation donnait-elle des indices pour discriminer des connaissances de type ACTIONS des connaissances de type OBJETS ? L'étude montrait que la polysémie concret / abstrait était une propriété générale de la nominalisation, cette dernière ayant pour fonction d'indifférencier ces oppositions. Par conséquent, si l'on voulait introduire cette distinction, il fallait coder ces traits dans la terminologie du domaine et non pas dans les propriétés linguistiques attachées aux opérations morphologiques. En effet, le fait que tout un chacun interprète *le garage de la voiture* plutôt comme un lieu (donc un OBJET) et non comme « l'action de garer son véhicule » est un cas de lexicalisation sémantique. La réponse que nous proposions à la suite de cette étude sémantique avait donc des conséquences sur l'organisation des connaissances, ce qui se traduisait en termes d'ingénierie linguistique par la création de dictionnaires spécialisés ou généraux (Clavier, 1999).

La seconde étude s'intéressait à une série morphologique limitée et devenue improductive, les noms dénominaux en –*age (branchage, cordage, personnage)* alors que les noms déverbaux en –*age* (*arrivage, arrimage)* se révélaient particulièrement productifs. L'originalité de l'étude était d'une part d'avoir dissocié les notions de productivité et de régularité souvent invoquées de conserve pour analyser des séries dérivationnelles ; en ce sens, la série faisait figure « *d'idiosyncrasie lexicale* », alors qu'elle manifestait une forte régularité. D'autre part, la « *très forte cohérence sémantique de la série* », amenait à s'interroger sur des phénomènes de « *captage* » entre les deux homonymes et permettait parfois de douter de l'existence de deux séries distinctes. L'étude faisait également des préconisations sur l'outillage théorique et formel à mobiliser pour mener à bien une étude sémantique. Elle mettait notamment en garde sur le rôle de la paraphrase pour décrire le sens :

> *Par ailleurs, on évitera de confondre paraphrase et modèle du sens. Si la paraphrase est une opération méta-linguistique indispensable pour appréhender intuitivement les significations, il*

> *ne s'ensuit pas qu'elle les exprime dans les catégories descriptivement et théoriquement les plus adéquates. Un modèle du sens digne de ce nom devrait prendre la forme d'un ensemble de primitives et d'opérations formellement définies, plutôt que celle de reformulations ad hoc en langue naturelle. (Berrendonner et Clavier, 1997 : p. 37)*

La proposition de modélisation s'écartait notablement des approches référentielles en vigueur – sous l'influence, notamment, du linguiste Georges Kleiber – et requérait pour sa formalisation des concepts puisés du côté des primitives cognitives (cf. individus discrets, classes, continuums et collectifs). Parmi les principales conclusions de l'article, figurait une conception « stratifiée » du lexique dans lequel plusieurs sous-systèmes se côtoient, dont certains, comme les noms dénominaux en *–age* peuvent s'éteindre en raison de la productivité de séries homonymes, ces derniers pouvant en partie remplir les mêmes fonctions sémantiques. L'étude se terminait par un ensemble de propositions expliquant pourquoi une série peut abonder en mots impossibles, plutôt qu'en mots possibles.

1.2.3.4. Morphologie dérivationnelle et analyse automatique de discours

Une seconde application de la morphologie dérivationnelle avait été envisagée en collaboration avec Geneviève Lallich-Boidin, Jacques Rouault et Ismaïl Timimi (Clavier et al., 1995). Isamaïl Timimi, à l'époque doctorant, mathématicien de formation, a soutenu une thèse en 1999[68] sur une méthode de classification destinée à extraire des énoncés entretenant des relations de paraphrase (Timimi, 1998). L'intérêt de cet article commun était double : d'une part, il permettait de montrer comment conjuguer des méthodes linguistiques et statistiques pour la recherche d'information ; d'autre part, il indiquait comment la normalisation des chaînes de caractères sur des critères morphologiques – flexionnels et dérivationnels – permettait de rapprocher des unités lexicales liées entre elles par des relations morphologiques, et plus largement de regrouper des énoncés en situation paraphrastique. Par exemple, les énoncés *interdiction de stationner sur les pelouses, stationnement sur les pelouses interdit* et *il est interdit de stationner sur les pelouses*

[68] Timimi Ismaïl, *De la paraphrase linguistique à la recherche d'information, le système 3 AD : théorie et implantation (aide à l'analyse automatique du discours)*, Thèse de doctorat en informatique et communication, sous la direction de Jacques Rouault, Université Grenoble 3, 1999.

(Clavier, 1996), ramenés à leurs formes canoniques respectives dans le cadre de la proposition pouvaient être considérés comme des reformulations d'un même contenu au sens de (Fuchs, 1994).

Le premier apport de la méthode résidait dans l'imbrication de méthodes d'analyse linguistique et de calcul de distance tels que Michel Pêcheux l'avait proposé en 1969 dans son ouvrage sur l'analyse automatique du discours (AAD), issu des travaux du linguiste Zellig Harris (1963). Cette méthode, aujourd'hui présentée dans les manuels d'analyse de discours (Maingueneau, 1991), s'inscrivait dans un cadre épistémologique qui allait au-delà de la simple entreprise technique de développement d'un instrument d'analyse ; l'AAD était une partie constitutive d'un projet de fondation d'une « psychologie sociale scientifique », c'est-à-dire un projet visant à identifier les traces socio-historiques révélatrices d'une idéologie dans les discours (Helsloot et Hak, 2000)[69]. Sur le plan technique, la méthode qui permettait de regrouper les énoncés proches dans des classes d'équivalence, s'appliquait à des corpus composés de discours produits dans des conditions homogènes ce qui assurait une certaine répétition dans le vocabulaire. L'analyse automatique des discours pouvait alors être considérée comme un outil de recherche des paraphrases d'un corpus[70].

Le principe de la méthode exposé dans (Timimi et Rouault, 1997 : p 6sqq), supposait de travailler sur des représentations des textes découpés en énoncés élémentaires ordonnés et de dimension fixe : huit places dédiées à chacune des catégories

[69] *« Pêcheux pose, […] la question de la place théorique du "discours" au sein du modèle saussurien. Le problème est le suivant : des interrogations telles que « Que veut dire ce texte ? » sont systématiquement exclues de l'analyse linguistique. Les réponses sont alors présupposées, abandonnées ainsi aux évidences de l'expérience empirique. Selon Pêcheux, c'est précisément le fait de « laisser à découvert le terrain » sans qu'il soit réinvesti par une autre science, qui autorise les idéologies à le (ré)envahir. En d'autres termes, bien que la linguistique se soit constituée en tant que science à travers une coupure épistémologique « saussurienne », elle a « oublié » de développer une théorie adéquate de la production du sens dans le discours. »* (Helsloot et Hak, 2000 : p. 14)

[70] Jacqueline Léon note que « *l'AAD69, conçu pour l'étude des processus discursifs et comme méthode alternative à l'analyse de contenu, se trouve à la croisée de la formalisation et de l'automatisation des sciences du comportement et de la linguistique, en plein essor dans les années 1960. Il va emprunter certains aspects de son automatisation aux deux grands domaines d'application non numérique des ordinateurs d'après la seconde guerre, la traduction automatique et la documentation automatique*. » (Léon, 2010)

grammaticales. Cette décomposition constituait « l'unité minimale d'assertion » et restituait le rapport sujet/prédicat tel que le définissait Antoine Culioli dans la notion de lexis. Ainsi, chaque énoncé était analysé sur le plan morphologique flexionnel et dérivationnel et était présenté sous la forme d'un couple *(f/t), f* étant la forme de la base lexicale et *t* ses traits morphologiques : catégorie grammaticale et variables descriptives. Les auteurs de cet article mentionnaient que l'on pouvait, à ce niveau, pondérer les traits morphologiques suivant le poids que représentait une catégorie en terme « d'apport informatif », les verbes et les noms étant plus informatifs que les déterminants ou les prépositions. La mise en œuvre de la notion de distance reposait à la fois sur des notions mathématiques et linguistiques.

Le calcul des distances entre énoncés était présenté comme une fonction de coût qui associait un nombre réel positif ou nul à une opération de transformation qui pouvait être une opération d'effacement, d'insertion ou de substitution. Ismaïl Timimi avait imaginé appliquer ces opérations à des mots pour rendre compte d'énoncés parasynonymiques ou antonymiques *: je préfère la campagne / j'aime la campagne / je choisis la campagne / je déteste la campagne* (Timimi, 1998). Dans (Clavier et al., 1995), il s'agissait de rendre compte de la distance qui séparait un mot construit de sa base lexicale en associant un coût à chaque opération dérivationnelle. Ainsi, comme la distance qui sépare le mot construit *nationalisation* de sa base *nation* est plus importante que la distance qui sépare le mot *parleur* de la base *parler,* le coût de transformation est plus fort dans le premier cas. En effet, alors qu'il faut quatre opérations dérivationnelles pour retrouver la base lexicale *nation*[71], il n'en faut qu'une pour retrouver la base *parler*. Suivant ce principe, les énoncés sont d'autant plus proches qu'ils mettent en jeu peu d'opérations dérivationnelles.

Cette méthode était intéressante dans la mesure où la notion de distance mathématique était associée à des opérations linguistiques et pas seulement à des éléments lexicaux, dont « l'apport informatif » nous semblait difficile à quantifier. Les limites de cette étude sont liées à la compositionnalité du sens : le caractère calculable et prédictible de chaque opération dérivationnelle n'est pas forcément

[71] ((((nation$_N$)-al)$_{Adj}$)-iser$_V$)-ation$_N$

assuré en raison des phénomènes de lexicalisation. Mais ceci est une limite propre à la modélisation du sens plus qu'à l'outil de recherche de paraphrases. Dans la littérature, il existe d'autres travaux qui ont associé la notion de distance mathématique à la morphologie pour améliorer la recherche d'information. Par exemple, Hongyan Jing et Evelyne Tzoukermann (1999), dans la continuité des travaux de Gérard Salton sur le modèle vectoriel, ont cherché à mesurer la distance entre des mots en leur associant les mots qui co-occurrent avec eux dans une fenêtre contextuelle (un contexte vectoriel comporte 10 mots). Les auteurs ont proposé d'élargir les mots recherchés à toutes les variantes morphologiques fournies par la base de données morphologique CELEX développée par le Max Planck Institute for Psycholinguistics. D'après les auteurs, cette méthode permet d'augmenter significativement le taux de précision. Alors que la méthode développée par Ismail Timimi permet de constituer des classes d'équivalence d'énoncés proches sémantiquement, la seconde méthode permet d'apprécier la proximité sémantique de deux mots en comparant la distance entre leurs vecteurs contextuels.

1.3. Bilan critique

1.3.1. Les apports à l'organisation des connaissances

En conclusion, cette première série de travaux s'inscrit dans le courant technique de *l'information retrieval* et développe une réflexion sur le langage et ses applications pour la recherche d'information, en particulier l'indexation automatique et l'extraction de connaissances. L'on peut résumer notre production scientifique comme suit : après avoir proposé un modèle de la dérivation suffixale pour représenter les connaissances linguistiques dans les textes intégraux (Clavier, 1996a), nous avons proposé une grammaire dérivationnelle et un modèle de dictionnaire pour permettre la conception d'un analyseur morphologique incluant la flexion et la dérivation (Clavier et Lallich-Boidin, 1994) (Clavier et Coret, 1997). Nous avons soumis notre modèle à une évaluation en corpus dans le cadre d'une campagne nationale d'évaluation (Clavier, 1996b). Nous avons participé à une réflexion portant sur l'intégration des modules de reconnaissance de la morphologie dans une

architecture multi-agents (Clavier et *al.*, 1996)[72]. Enfin, nous avons proposé une contribution de l'analyse morphologique à l'extraction de connaissances dans un manuel d'utilisation (Clavier et al., 1997) (Clavier, 1999), ce qui suppose, en sus d'une reconnaissance des constituants internes aux mots construits, une analyse sémantique. Deux études ont été conduites en ce sens, l'une sur les phénomènes de concurrence suffixale, qui sont des cas particuliers de la synonymie (Clavier et Coret 1998), l'autre sur l'homonymie suffixale (Berrendonner et Clavier, 1997).

De notre point de vue, l'apport essentiel de nos travaux réside dans le choix des morphèmes comme unité de représentation des connaissances textuelles. Ce choix n'a rien de novateur. Bien avant la théorie de l'information qui considère les formes graphiques comme des unités de code, le mot était considéré comme l'unité porteuse de signification jusqu'au XVIIIème siècle. Mais d'après Oswald Ducrot et Tzvetan Todorov, c'est *« l'avènement de la linguistique comparative* [qui] *a imposé une dissociation du mot en unités significatives complémentaires […] la comparaison de deux langues différentes en vue d'établir leur parenté ne* [pouvant] *se faire mot à mot mais de partir de mot à partie de mot. »* (Ducrot et Todorov, 1972 : p. 257). Dès le structuralisme, bien que Saussure le considérât comme incontournable, tous les courants linguistiques ont cherché à évacuer le mot que ce soit la glossématique, le fonctionnalisme ou le distributionnalisme… Cependant, ainsi que l'indique Jacqueline Léon (2001), « *la traduction automatique a ravivé cette question du mot en lui donnant un nouvel enjeu, celui de définir des unités pour la machine qui soient aussi des unités linguistiques et des unités de traduction. Il s'agissait de faire coïncider forme graphique, unité syntaxique et unité sémantique* ». De manière symétrique, la recherche d'information pose aussi la question de l'adéquation entre des mots, ceux des requêtes formulées par les utilisateurs et ceux des documents.

La problématique de mise en correspondance formelle entre des mots a été le moteur de nos recherches en morphologie. Une grande partie de notre travail de thèse a été consacrée à la description des unités formelles internes aux mots, au choix des formes

[72] Clavier V., Warren K., Lallich-Boidin G. et Stéfanini M.-H. (1996), « Intégration de la morphologie dérivationnelle dans un système distribué d'analyse du français écrit », *in Actes du colloque "Informatique & Langue Naturelle" (ILN'96)*, 9-10 octobre 1996, Université de Nantes, p. 103-120.

canoniques à la justification des alternances formelles, à la structuration des niveaux d'information. Ces analyses se sont traduites concrètement par la description de 8000 lexèmes. Nous pensions alors que l'un des apports du TAL n'était pas seulement le développement de formalismes logiques, aspect qui fait totalement défaut dans notre travail, mais bien une contribution importante à la constitution de ressources lexicales ainsi que cela se pratique dans le domaine de la dictionnairique et de l'ingénierie linguistique. En 2005, Nabil Hathout faisait malheureusement le constat qu'il n'existe encore qu'un seul dictionnaire morphologique de langue française[73] à partir duquel l'on peut extraire des connaissances morphologiques « sûres » pour le français[74]. Il s'agit du *Dictionnaire Synchronique des Familles de Mots* (DISFA) constitué par Claude Gruaz et auquel nous avons travaillé pendant un an[75]. La réalisation de ces ressources nous avait donné une rigueur dans la description des données langagières qui pouvait être réinvestie dans divers projets d'ingénierie linguistique[76].

Pour terminer, nos travaux offre une contribution à l'organisation des connaissances pour les raisons suivantes :

- **La question de l'organisation des connaissances est posée dans un contexte précis** : celui de la recherche d'information entendue comme l'ensemble des techniques et méthodes issues du traitement automatique du langage permettant de représenter l'information textuelle dans un système de reconnaissance automatique du français.

[73] Hormis la base de données Celex développée par le Max Planck Institute für Psycholinguistics http://celex.mpi.nl/scripts/entry.pl

[74]Hathout Nabil (2005) «Acquisition de connaissances morphologiques à partir de dictionnaires », http://w3.erss.univ-tlse2.fr/textes/operations/operationTAL/UE_TAL/2005-2006/LexicoMorpho/morpho-dico-hathout.pdf.

[75] sans être citée d'ailleurs : http://sites.univ-provence.fr/delic/disfa/accueil.html

[76] Au début des années 90, de nombreux projets de bases de données lexicales (Pérennou et al., 1992) ou de dictionnaires électroniques (Courtois et Silberztein, 1989) voient le jour.

- **L'origine des connaissances est identifiée :** les connaissances émanent des documents textuels eux-mêmes et non pas de sources extérieures aux documents (classifications, thésaurus, ontologies, etc.)
- **Le principe d'organisation des connaissances est explicité :** il s'appuie sur les niveaux de description du langage et engage des choix linguistiques sur les unités du langage à décrire (les morphèmes), leur forme (les bases régulières, alternantes, courtes ou longues, les suffixes) ; leur catégorisation (syntaxique), leur description sémantique (en termes de combinatoire catégorielle) et les relations qui existent entre les membres d'une série dérivationnelle.
- **Le choix de la nature des connaissances à décrire est justifié :** les connaissances à décrire sont d'ordre morphologique, puisque la productivité lexicale repose en grande partie sur le recours aux procédés de construction des mots. Or, si les systèmes de reconnaissance automatique n'ont pas une couverture lexicale suffisante, il y a échec de la reconnaissance, ce qui induit des problèmes de silence.
- **Les outils de l'organisation des connaissances sont décrits :** la mise en œuvre d'un système de reconnaissance automatique pour reconnaître les mots construits par suffixation a conduit à définir les dictionnaires, les principes de structuration des entrées, la nature des informations à y faire figurer ainsi que les grammaires, notamment les règles de régularisation des alternances formelles.
- **L'environnement dans lequel s'insèrent ces outils est pris en compte :** la réflexion sur le module morphologique dérivationnel ne s'est pas conduite *ex nihilo* mais s'est articulée avec les travaux conduits par les informaticiens sur une architecture multi-agents intégrant tous les outils déjà existants dans le laboratoire, notamment l'analyseur morphologique flexionnel, le dictionnaire des bases et l'analyseur syntaxique qui était en cours d'élaboration.
- **La finalité de l'organisation des connaissances est envisagée :** les connaissances recueillies sur la morphologie dérivationnelle permettent de structurer les index automatiques, de les réduire et de les documenter avec des informations linguistiques. Ces informations permettent de définir un

niveau de représentation commun aux requêtes et documents – ce qui favorise leur rapprochement – et homogène du point de vue des unités qui y figurent.

- **Une application dans le contexte professionnel est abordée :** dans les entreprises, la mise à disposition de la documentation technique sur support électronique se généralise. C'est notamment le cas des manuels d'utilisation de logiciels en informatique. La mise en oeuvre de méthodes d'extraction de connaissances permet à des utilisateurs experts en informatique de rechercher des informations fines leur permettant de mener à bien certaines tâches : dépannage, maintenance, installation. Dans ce contexte, les informations de nature sémantique associées aux connaissances morphologiques sont nécessaires : des propositions sur la contribution de la morphologie dérivationnelle à l'extraction de connaissances de type ACTION versus OBJET ont été faites et se sont traduites en termes d'ingénierie linguistique.

1.3.2. Les limites

Les limites sont de plusieurs ordres, et souvent, un travail de doctorat pluridisciplinaire consiste à l'évaluer à l'aune de chaque discipline concernée, ce que l'on peut faire rapidement bien que cela ne soit pas en notre faveur.

Concernant les aspects descriptifs en linguistique, le nombre impressionnant de suffixes abordés rendait l'étude superficielle. Il n'est pas rare de voir des linguistes traiter un seul suffixe[77]. Par conséquent notre travail souffre de manques comme par exemple, la non prise en compte des lacunes lexicales qui sont sans conteste plus instructives que l'étude des néologismes sur le plan linguistique ; un traitement superficiel de la question de la productivité lexicale ; l'absence de réponse sur le nombre de suffixes en français ou encore une description de la composante sémantique insuffisante.

[77] Dal Georgette (1994), *Un exemple de traitement associatif du lexique construit : analyse unificatrice des mots suffixés par -et(te),* Thèse de doctorat sous la direction de Danielle Corbin, Université de Lille III, 354 pages.

Concernant la dimension formelle de la linguistique, la question du choix des modèles théoriques de référence pour traiter les règles est seulement esquissée et leur choix engageait les rapports entre morphologie et syntaxe. Bien que nous ayons évoqué la question d'une formalisation de la dérivation sous la forme d'opérateurs/opérandes, la description est loin d'être menée à son terme. Elle n'est pas suffisamment étayée par exemple lorsqu'il s'agit de décrire l'enchaînement récursif de plusieurs opérations d'affixation (préfixe et suffixe).

Au sujet des aspects relevant de la linguistique appliquée, nous avions fait le choix de nous intéresser aux ressources, mais nous n'avons ni créé de dictionnaire électronique complet à mettre à la disposition de la communauté en TAL, ni conçu d'analyseur opérationnel. Ainsi que nous l'avons indiqué, la plupart des ressources dictionnairiques pour le TAL émanaient du LADL. Cependant, à la grande différence de notre laboratoire, le modèle de description était donné et le travail des linguistes consistait à recenser en grandeur réelle les structures syntaxiques et morphologiques élémentaires suivant le modèle théorique retenu. Quant à l'implémentation, les cours de Prolog que nous avons reçus dans le cadre de notre formation, nous ont permis de faire des mini-programmes, mais pas une implémentation grandeur réelle.

Enfin, les méthodes utilisées se limitaient aux méthodes distributionnelles de segmentation qui consistent à mettre au jour les rapports formels que les mots entretiennent entre eux en respectant le principe de compositionnalité. Toutefois, on peut déplorer que notre analyse ne se déroulât pas en synchronie puisque le recueil de mots sur lequel nous travaillions était issu de dictionnaires (le DELAS, lui-même issu du Petit Robert). Or, les dictionnaires intègrent les néologismes à retardement et conservent les vestiges lexicaux. Le recours aux dictionnaires de langue se révélait cependant indispensable pour guider notre intuition en particulier pour les mots dits « rares » ou étymologiquement opaques. L'introduction de corpus hétérogènes dans les procédures d'évaluation, sans aucun critère de sélection était également critiquable et l'on peut même se demander comment l'on pouvait accorder du crédit aux résultats de l'évaluation des analyseurs : la langue « standard » n'était pas la somme des vocabulaires des corpus retenus…

Pour terminer sur le TAL, les spécialistes eux-mêmes reconnaissent que l'on a dû en rabattre sur la « théorie », et que son succès réside dans la « raison pratique » pour reprendre les termes de David Piotrowski (1999)[78]. Quant à ses apports pour la recherche d'information, Mathieu Valette et Monique Slodzian posent la question de savoir si le TAL n'aurait pas *« construit sa prospérité sur ce point aveugle qu'est l'appariement de mots »* (Valette et Slodzian, 2008 : p. 120). Les auteurs décrivent alors les impasses dans lesquelles le TAL s'est engagé. D'abord, l'on a cherché à enrichir l'indexation de documents au moyen d'analyseurs morphosyntaxiques à large couverture, en cherchant à améliorer les performances des mots-clés par le recours à la racinisation ou à la lemmatisation. Afin de limiter le bruit, l'on a fait appel à la syntaxe pour désambiguïser les mots ; de nombreuses études ont été réalisées sur les termes complexes qui ont abouti à la création d'outils d'extraction sophistiqués. Et… les auteurs de conclure que les résultats sont assez décevants :

> *« Au fil des campagnes d'évaluation successives, il est apparu que l'addition systématique de ressources linguistiques de bas niveau (morphologiques et syntaxiques) n'améliorait pas automatiquement les résultats et, en particulier, que les systèmes à base de morphosyntaxe s'avéraient décevants (Jacquemin 2000). On doit aux travaux menés sur les techniques et outils d'évaluation d'avoir permis ces clarifications en montrant la complexité des interactions entre faits linguistiques et d'avoir ainsi attiré l'attention sur les limites d'une conception logiciste de la « langue naturelle » (Poibeau 2003). » (Valette et Slodzian, 2008 : p. 121).*

En ce qui concerne plus particulièrement la morphologie, il est certain que les traitements automatiques ont un coût. Dans un rapport datant de 2005 consacré à l'évaluation de la pertinence des résultats d'un SRI utilisant un lemmatiseur et/ou un raciniseur et/ou un analyseur morphologique, Fabienne Moreau et Pascale Sébillot présentent les résultats suivants (Moreau et Sébillot, 2005 : p. 11):

- La lemmatisation est dans 40% des cas l'une des meilleures méthodes pour retrouver et classer des documents (expérience conduite par Hull et Grefenstette en 1996) ; les cas d'échec sont dus aux mots inconnus. Selon

[78] Piotrowski D. (1999), « Le T.A.L.N. Faillite des ambitions théoriques, Succès de la raison pratique » in *Sémiotiques*, n°17, p. 1-27.

ces auteurs, pour le français, les résultats seraient meilleurs à condition d'utiliser un lemmatiseur non tributaire d'un dictionnaire.

- La racinisation avec des méthodes algorithmiques de type Porter produit un nombre d'erreurs trop important pour le français et montrent des limites certaines ;
- L'apport de l'analyse dérivationnelle semble moins net, les auteurs mentionnant qu'elle présente l'avantage de retrouver plus de variantes. Dans ce cas, il faut une procédure de contrôle qui valide sémantiquement les variantes morphologiques ;

Signalons en revanche, que l'extension de requêtes a connu un meilleur sort et qu'elle a fait l'objet de nombreux travaux. Par exemple Fabienne Moreau et Vincent Claveau proposent d'enrichir les requêtes en acquérant automatiquement des relations morphologiques sans qu'il y ait de connaissances externes (règles de réécriture, bases de suffixes, lexiques) (Moreau et Claveau, 2006). Cette étude est très révélatrice du nouveau tournant qui s'est imposé notamment avec l'accès généralisé au web. Pour les auteurs, les nouveaux défis auxquels doivent faire face les chercheurs sont d'augmenter la portabilité des outils, afin qu'ils ne soient liés à aucune ressource, à aucune langue, à aucune intervention humaine. **Ainsi, s'ouvrait l'ère des méthodes et techniques sans connaissances, pour lesquelles il n'est pas nécessaire de savoir ce qu'est un mot construit.** Les corpus d'entrainement, les méthodes d'apprentissage permettent de détecter les variantes morphologiques « avec une très bonne couverture et une grande précision » (*ibid.*)

1.3.3. Pour conclure sur l'apport de la morphologie à la recherche d'information

La perspective de l'indexation que nous avons retenue est, comme l'indique Muriel Amar, « instrumentale » (Amar, 2000 : p. 26-32), puisqu'elle n'est pas envisagée comme un processus d'interprétation humaine du contenu d'un texte, mais bien comme un processus de médiation qui imbrique étroitement les connaissances linguistiques et la technique. Le résultat de cette indexation peut être évalué en regard de critères de mesure propres à la recherche documentaire :

- les traitements sont définis pour limiter le silence, i.e améliorer le taux de rappel : le silence concerne les séries dérivationnelles dans lesquelles les variations de formes sont importantes, comme les cas de supplétion lexicale ex. *foie / hépatite /hépatique*) ;
- les traitements sont également définis pour limiter le bruit, améliorer le taux de précision. Le bruit renvoie à tous les cas d'ambiguïtés de la langue naturelle, comme l'homonymie catégorielle (*plumage* dérivé de *plume* ou de *plumer*), ou l'homonymie suffixale (*–ment* est un suffixe dérivé d'un adjectif *lent* / *lentement* ou d'un verbe *ranger / rangement*).

L'apport de la morphologie en reconnaissance se situe à deux niveaux. Au premier niveau, elle permet de délimiter sur des critères formels les frontières entre les cas de dérivation morphologique, avec une régularité formelle justifiable sur le plan phonologique, et les cas de synonymie lexicale, sans lien formel mais uniquement sémantique. Au second niveau, elle permet d'augmenter le nombre d'homonymes sur des critères morphologiques : par exemple, la possibilité de distinguer plusieurs suffixes de noms homonymes en *–age*. La finesse de la description va malheureusement de pair avec l'augmentation du nombre d'ambiguïtés.

2. Les textes comme modèle d'organisation des connaissances : *profiler, annoter, classer*

Les notions clés de l'organisation des connaissances qui sont envisagées dans ce chapitre prennent place dans le contexte de l'internet. Deux conséquences en ont résulté : d'une part, la mise à disposition des chercheurs de très grands volumes de données et, d'autre part, la nécessité de recourir à des méthodes de traitement de l'information plus robustes, comme la fouille de textes et la linguistique de corpus. Notre cadre de réflexion s'inscrit dans un nouvel environnement technique, celui du développement de méthodes de classification non supervisée pour classer des mots ou des textes.

La classification automatique comporte une étape d'organisation des données textuelles qui nécessite d'abord d'indexer les documents – il faut trouver des descripteurs qui caractérisent au mieux les données – et ensuite, de comparer la distribution des index de plusieurs documents afin de les regrouper suivant des critères de proximité. Ainsi, plutôt que d'envisager les textes comme des « sacs de mots », nous avons fait des propositions pour prendre en compte les textes annotés au moyen de variables morphosyntaxiques, ce qui permet d'appréhender le genre textuel. Les méthodes de classification peuvent alors permettre de classer des documents en genre et domaine, ce qui se révèle particulièrement intéressant pour la recherche d'information spécialisée.

Profiler, annoter et *classer* sont donc les nouveaux objectifs qui s'appliquent à des corpus choisis pour entraîner les données d'apprentissage. Le nombre de travaux qui portent sur ce thème est de 6 sur 45 entre 2002 et 2006 (quatre autres publications sont des retours d'expériences pédagogiques, ce qui aboutit à 10 publications réalisées entre 2001 et 2006)[79].

2.1. L'environnement de notre recherche : éléments de cadrage

2.1.1. La classification automatique et la recherche d'information

La classification automatique est issue des méthodes d'analyse de données visant à extraire des données, voire des connaissances, dans de grandes bases de données informatiques de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données[80]. Lorsque les données à classer s'appliquent à des textes, l'on parle de *text mining* (fouille de textes), et dans ce cas, les méthodes de traitements s'appliquent à du texte plat (ou du texte brut) c'est-à-dire des données qui ne comportent aucune information ajoutée, ni annotation, ni marque de structuration.

[79] Références citées : Ref. 21, 22, 23, 24, 28 et 29. Les autres références de la même époque sont liées à des retours d'expériences pédagogiques (Ref. 20, 25, 26 et 27).

[80] (Tufféry, 2010 : 4) cité par (Macmurray, 2012 : p. 40).

Les textes eux-mêmes sont stockés dans des bases de données thématiques, les applications les plus fréquentes sont l'extraction d'informations pour la veille et la recherche d'associations[81].

Les méthodes de classification automatique de textes entretiennent des liens étroits avec le modèle vectoriel proposé par Salton & Mc Gill en 1983, Salton étant l'un des pères fondateur de *l'information retrieval.* Le principe consiste *« à représenter des documents par des vecteurs calculés à partir des mots les plus significatifs présents dans chaque document. Ces vecteurs sont ensuite regroupés par similarité de manière à classer ensemble les documents traitant des thèmes similaires. Cette classification peut alors servir à l'indexation et à la recherche des documents, mais aussi à l'extraction d'informations plus élaborées.* » (Memmi, 2000). La dimension textuelle visée par la classification est le plus souvent la thématique, une préoccupation centrale de la recherche d'information.

Suivant Daniel Memmi, pour représenter des textes dans un espace vectoriel, il convient tout d'abord de choisir la partie ou les parties significative(s)du document pour établir les vecteurs : tout le texte, une partie du texte comme le résumé, l'introduction, les titres, etc. Plus le document est long, plus il faut le découper en petites unités pour les intégrer dans une matrice. Ensuite, il faut choisir les traits descriptifs des représentations vectorisées. Dans les premiers modèles, on retenait les mots pleins de fréquence moyenne. Pour les langues à morphologie riche, l'auteur mentionne qu'il est préférable de lemmatiser les formes graphiques, voire d'opérer des traitements morphologiques plus évolués pour ramener les lemmes à leur troncat, par troncature, ou à leur racine, par une analyse morphologique. Enfin, la troisième étape consiste à employer une méthode de classification pour regrouper les vecteurs similaires et faire émerger des proximités thématiques.

2.1.2. La linguistique de corpus et le TAL

A partir des années 90, l'essor de l'internet et des méthodes dites « stochastiques » (statistiques et probabilistes), a conduit à prendre massivement en compte les

[81] (Feldman & Sanger, 2007) cité par (Macmurray, 2012 : p. 42).

« vraies données langagières ». Le « TAL robuste » (Cori, 2008*)* ouvrait la voie de la linguistique de corpus. Outre les méthodes fondées sur des règles (méthodes symboliques) et les méthodes numériques, il existait les méthodes à base d'automates d'états finis ou d'expressions régulières, qui, selon Marcel Cori, constituaient une approche intermédiaire entre le « TAL théorique » et le « TAL robuste » permettant d'intégrer le contexte.

Le travail sur corpus établissait en effet une relation étroite avec le TAL, mais il ne s'agissait plus d'une conception théorique du TAL mais d'une conception empirique, plus tolérante aux erreurs, ce que l'on nomme la « robustesse », ou le « TAL robuste ». Marcel Cori explique que le déplacement des enjeux de recherche du TAL est dû à deux insuffisances du « TAL théorique » (Cori, 2008). D'une part, les grammaires et les lexiques ne peuvent être exhaustifs en raison de l'apparition quotidienne de nouveaux mots et des productions déviantes des locuteurs vis-à-vis de la norme graphique. D'autre part, les analyseurs syntaxiques produisent un nombre considérable d'ambiguïtés. C'est dans le sillage de ces critiques adressées au TAL théorique que s'est développé le TAL robuste qui, selon Marcel Cori, doit respecter trois critères (*ibid.,* 2008, p. 98) :

a) il faut que le logiciel prenne pour données de « vraies » productions langagières, du texte tout venant, et non des exemples forgés par le linguiste ;
b) les logiciels doivent fournir une solution et une seule, et le système ne doit pas se bloquer sous prétexte que les données sont « incorrectes » ou agrammaticales ;
c) enfin, les logiciels doivent se prêter à une évaluation quantitative de leurs performances.

2.1.3. Le LIFO et le CORAL : des cadres de recherche mono-disciplinaire

Le laboratoire d'informatique fondamentale d'Orléans (LIFO) est le premier laboratoire à nous avoir accueillie en 1998 en tant que membre permanent ; le centre orléanais de recherche en anthropologie et linguistique (CORAL) est le second. En l'absence d'un laboratoire en sciences de l'information et de la communication, nous

nous sommes intégrée dans ces laboratoires en faisant valoir nos centres d'intérêt pour la recherche d'information et le TAL.

Contrairement à la conception de la recherche d'information qui présidait au laboratoire pluridisciplinaire que nous avions connu à Grenoble, la recherche d'information à Orléans était envisagée comme une spécialité de l'informatique au sens défini par Mohand Boughanem : *« La recherche d'information (RI) […] est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information. »* (Boughanem, 2008 : p. 19). Au contact de cette conception très informatique de la recherche d'information, nous prenions alors pleinement conscience de notre décalage et n'étions pas en mesure de faire valoir nos compétences ni en linguistique ni en sciences de l'information et de la communication. En effet, bien que cette équipe ait développé des travaux sur la conception d'un système de recherche d'information dénommé Dialect (Braunwarth et al., 1994), il n'y avait aucun outil linguistique, dictionnaire, analyseur ou corpus de données textuelles. Il n'y avait pas non plus d'applications qui pouvaient justifier notre participation.

Rapidement, nous avons fait le choix de travailler avec d'autres chercheurs du LIFO qui souhaitaient appliquer leurs méthodes de classification à des données textuelles et qui étaient intéressés par une collaboration. Dans les années 2000, la problématique de la recherche d'information avait pris place dans un nouveau contexte fortement lié à la montée de l'internet, ce qui avait eu une double conséquence : d'une part, les données textuelles étaient devenues plus facilement accessibles et d'autre part, les méthodes numériques plus adaptées aux traitements de grands ensembles devenaient centrales. La linguistique de corpus et la fouille de texte avaient localement – mais aussi nationalement – éclipsé les recherches sur le TAL théorique. Cependant, les méthodes se révélaient différentes, les corpus étant considérés comme un réservoir de connaissances à exploiter pour l'indexation et pour la classification de textes. La perspective considérée était toujours la recherche de méthodes et de techniques pour la recherche d'information, mais ce n'étaient pas les mêmes communautés qui étaient sur le devant de la scène. La linguistique formelle s'effaçait au profit des approches empiriques de la linguistique de corpus, les

informaticiens issus du courant de la programmation logique laissaient la place aux apprentistes. Ce changement de paradigme était d'ailleurs assez douloureusement vécu au sein du laboratoire.

Dans cet environnement, tout était nouveau pour nous : les méthodes mathématiques, les statistiques, les probabilités, utilisées pour la formalisation ne nous étaient pas familiers et, parallèlement, nous devions apprendre à constituer des corpus, à les traiter, les annoter. Nous pouvons considérer cette période comme une mise à jour de nos compétences, alors même que nous commencions notre carrière d'enseignant-chercheur… Le problème central que nous rencontrions était l'absence d'environnement linguistique propre à la recherche d'information. Nous avons tenté d'y remédier en agissant sur deux plans : la formation et les projets de collaborations entre laboratoires.

En ce qui concerne la formation, nous sommes intervenue dans le DEA d'informatique et nous avons proposé des cours de linguistique pour le TAL ; à l'UFR de lettres, qui ne possédait pas de spécialité en TAL, nous avons proposé des cours de TAL en Deug et en licence, puis nous avons encadré des mémoires de maîtrise et de DEA sur ce sujet. Progressivement, une mention « traitement de l'information » a été introduite et nous avons créé notre premier cours de statistique textuelle en licence et maîtrise. C'est ainsi que nous avons fait la connaissance de Céline Poudat au CORAL et de Guillaume Cleuziou au LIFO dont nous avons encadré les travaux de recherche. La formation était le point central de nos efforts et nous avons publié un article commun avec Céline Poudat sur l'enseignement du TAL dans les cursus de lettres et de langues (Clavier et Poudat, 2001). Cette communication a été présentée au congrès international sur la traduction automatique : Machine translation Summit[82].

En ce qui concerne les collaborations scientifiques, Gabriel Bergounioux, à l'époque directeur du CORAL, nous avait confié la responsabilité d'un petit groupe de

[82] CLAVIER V., POUDAT C. (2001) « Teaching Machine translation in non computer science subjects : report of an educational experience within the University of Orléans » In *Actes du VIIIème congrès international de Traduction Automatique (Machine Translation Summit)*, du 22 au 28 septembre 2001, Saint-Jacques de Compostelle, Espagne, p. 19-23.

recherche dénommé « Contexte, concept et acculturation » qui rassemblait trois enseignants-chercheurs, quatre avec moi, et une doctorante (Céline Poudat). Avec Céline Poudat, nous avons organisé des séminaires de formation destinés aux chercheurs sur la linguistique de corpus, notamment sur l'encodage de corpus en XML, le balisage TEI ; nous avons rédigé notre premier contrat de recherche qui, certes n'a pas été retenu, mais qui a eu le mérite d'associer des chercheurs du LIFO et du CORAL. Enfin nous avons développé un projet de plate-forme linguistique destinée à accueillir des corpus– pour que les apprentistes puissent entraîner leurs données – et des outils – pour que les linguistes puissent traiter leurs corpus.

Au bout de quelques années d'efforts, nous sommes parvenue à ce que nous considérions comme un véritable succès : des publications scientifiques communes sur des méthodes de classification qui, au lieu de s'appliquer à des sacs de mots, s'appliquaient à des corpus sélectionnés, traités, annotés, et qui restituaient une dimension textuelle fondamentale : celle du genre. Ces méthodes ont été utilisées pour des applications qui relèvent de la recherche d'information. Ce travail s'est déroulé dans le cadre du co-encadrement des thèses de doctorat de Guillaume Cleuziou et de Céline Poudat.

2.2. L'apport de la linguistique à la classification

2.2.1. Classification, distance mathématique et données textuelles

2.2.1.1. Enjeux de la classification de textes

Le co-encadrement de la thèse de Guillaume Cleuziou en informatique, dont la préparation s'est déroulée entre 2001 et 2004, nous a amenée à travailler dans le cadre d'un modèle mathématique hérité du modèle vectoriel. L'objectif de la thèse était de développer des méthodes de classification et un chapitre en particulier était consacré à la recherche d'information. La classification permet d'analyser des volumes importants de données textuelles, cherche à construire des ensembles homogènes de textes, qui sont considérés comme des individus suivant la tradition statistique et qui partagent des caractéristiques communes. La classification peut être

supervisée ou non supervisée. Dans le premier cas, il s'agit d'apprendre à classer un nouvel individu – une page web, un document, un mot, etc. –parmi un ensemble de classes prédéfinies à partir de données d'entrainement (Cleuziou, 2004 : p.7). Dans le second cas, la classification non supervisée consiste à extraire des classes ou groupes d'individus à partir d'une population présentant des caractéristiques communes, le nombre et la définition des classes n'étant pas donnés *a priori* (*ibid.* : p.7). La classification relève des techniques d'analyse de données multidimensionnelles, du domaine de l'intelligence artificielle, et plus particulièrement de la reconnaissance des formes.

Les méthodes peuvent se distinguer suivant qu'elles permettent de classer un individu dans une seule classe, c'est le « *hard clustering* » ou dans plusieurs classes, le « *soft clustering* ». Il existe enfin plusieurs méthodes de classification et nous renvoyons à l'ouvrage de Fidelia Ibewkwe-San-Juan pour un exposé détaillé et plus accessible à un lecteur des sciences de l'information et de la communication (Ibekwe-SanJuan, 2007 : p. 59-88). Ainsi que le présente cet auteur, il existe les méthodes dites hiérarchiques qui construisent une hiérarchie de classes emboîtées en utilisant des mesures de similarité, l'information mutuelle par exemple. Il existe les méthodes de partitionnement, comme la méthode des *k-means* qui permettent de choisir le nombre de classes et qui s'appuient sur des mesures de distance euclidienne. Il existe enfin des méthodes issues de l'analyse de données, comme l'analyse factorielle des correspondances qui emploie la distance du χ^2 ou l'analyse latente sémantique qui est utilisée par exemple, pour identifier des synonymes.

Lorsque les individus à classer sont concrets, comme des champignons par exemple, les critères de classification sont objectivables : la couleur, la forme, la texture, etc. En revanche, lorsqu'il s'agit de classer des textes, les critères sont moins aisés à trouver, précisément en raison des différents niveaux de caractérisation des textes qui se projettent sur les formes. Or, des textes peuvent être comparables en raison de leur style[83], des genres auxquels ils appartiennent, (roman, poésie, théâtre), du sujet, du domaine (droit, médecine, littérature) de la longueur des textes (roman, nouvelles),

[83] La stylométrie est un domaine important des statistiques textuelles qui peut permettre de rechercher la paternité d'une œuvre (cf. la paternité des œuvres de Corneille et Molière).

etc. Par conséquent le premier travail consiste à définir le niveau de description des données et mettre en œuvre des critères de constitution de corpus : il faut comparer ce qui est comparable. La classification se révèle particulièrement intéressante pour la recherche d'information, puisque ces méthodes reposent sur l'hypothèse que des documents de contenus similaires seront pertinents pour les mêmes requêtes (Van Rijsbergen, 1979). Pour Guillaume Cleuziou, il s'agissait fondamentalement d'un « *problème d'organisation des données textuelles* » préalable à l'indexation des documents (*ibid.,* p. 3), ce qui nous ramenait à notre préoccupation initiale concernant la nature des représentations textuelles sur lesquelles s'appliquent les traitements de l'information.

Par ailleurs la classification soulevait un problème mathématique relatif à l'établissement de « mesures de similarité » qui nécessitaient de « connaître et quantifier les relations » existant entre les représentations des textes. Guillaume Cleuziou a utilisé –et a développé également– diverses mesures qui s'appliquent à plusieurs types de représentations textuelles. Dans le chapitre 5 de sa thèse de doctorat, l'auteur recense les mesures existantes qu'il a utilisées (*ibid.* p. 129sqq) :

a) Les mesures statistiques de cooccurrences permettent de découvrir les récurrences lexicales. (*ibid.,* p. 129) ; les mesures les plus courantes sont la mesure d'information mutuelle, la méthode des mots associés ; le coefficient de Dice et la mesure de Jaccard. Elles sont utilisées principalement pour l'extension de requêtes.
b) Il existe également les mesures fondées sur des distributions. Ces méthodes se fondent sur l'hypothèse harrissienne suivant laquelle « *les mots qui apparaissant dans des contextes similaires tendent à avoir des sens similaires* ». Alors que les méthodes de cooccurrences ne prennent en compte que les relations de proximité, les méthodes distributionnelles recherchent des mots apparaissant dans des contextes syntaxiques similaires. Ces méthodes permettent la recherche des synonymes d'un mot, qui, en principe,

n'apparaissent pas tous ensemble au sein de la même « fenêtre contextuelle »[84].

c) Une troisième famille d'approches pour évaluer la proximité entre les mots, consiste à intégrer des connaissances syntaxiques et sémantiques pour définir les mesures. L'auteur indique que ces connaissances sont alors formalisées par une taxinomie de concepts et précise que « *la principale limite de ce type d'approche est que, le plus souvent, il n'existe pas de base de connaissances spécifique à un domaine suffisamment structuré* ».

Nous avons produit quatre publications avec Guillaume Cleuziou et Lionel Martin, second co-encadrant de la thèse pour les aspects informatiques : (Clavier et *al*. 2002 ; Cleuziou et *al*. 2003 ; Cleuziou et *al*. 2004a et 2004b). Ces travaux sont issus de longues séances de réunions, au cours desquelles nous cherchions à trouver un vocabulaire commun, à définir les notions de distance sur le plan mathématique et sur le plan linguistique ; sur la manière de caractériser un texte, la plupart des travaux de fouille de textes ne connaissant que les « sacs de mots ». Dans les publications collectives produites, notre contribution portait plus particulièrement sur la caractérisation des données textuelles ainsi que sur la définition d'applications possibles pour la recherche d'information.

2.2.1.2. Classification de mots et profils utilisateurs

La première publication collective développait des méthodes de classification non supervisée destinées à organiser des mots en un profil utilisateur (Clavier et *al*., 2002), cette étape constituait une « *phase de présélection d'un ensemble de données pouvant conduire à l'acquisition de connaissances* » (*ibid.,* p.226) et s'inscrivait dans le contexte général de l'organisation des connaissances sur le web.

La notion de « profil utilisateur » est utilisée couramment en documentation, en particulier dans la veille. Elle avait été remobilisée à la fin des années 90 par Sylvie Lainé-Cruzel et Christine Michel dans une série de travaux portant sur un système de recherche d'information piloté par les profils dénommé Profil-Doc (Lainé-Cruzel,

[84] Les mots qui apparaissent ensemble dans les mêmes phrases ou séparés par un nombre limité de mots.

1999 ; Michel et Lainé-Cruzel, 1999). Dans le système Profil-Doc, chaque utilisateur devait, préalablement à sa recherche, fournir des renseignements sur lui-même (par exemple son niveau de formation), sur les thèmes et sur ces objectifs de recherche. Le profil consistait alors en un ensemble de critères documentaires qui étaient utilisés pour « *filtrer une information exploitable* » dans des bases de données scientifiques. Dans un environnement ouvert tel qu'internet, la notion de profil devait être redéfinie. Certains moteurs de recherche faisaient usage de la notion de profil « a posteriori *pour filtrer l'information, la visualiser et parfois pour étendre les requêtes* » (Clavier et *al*., 2002 : p. 220-221). Nous évoquions alors le moteur de recherche Vivisimo qui proposait de classer les documents sous forme de listes de thèmes hiérarchisés ce qui offrait l'avantage d'« *optimiser la recherche d'information, […] le principe d'une présentation thématique de l'information par rapport à un affichage par score de pertinence permettait à l'utilisateur de visualiser la manière dont les connaissances sont organisées sur le web et donc de se les approprier* » (*ibid.* : p. 221). Un autre exemple d'application du profil utilisateur consistait à étendre les thèmes de la requête. Pour cela, deux méthodes étaient possibles, l'une mise en œuvre par Vivisimo qui recourait à des techniques de classification à partir des titres, des URL ou de courtes descriptions dans des domaines ciblés ; l'autre mise en œuvre par LivesTopic qui faisait appel à des taxinomies. Notre appréciation des méthodes était en la faveur de Vivisimo et nous indiquions que « *la constitution de groupes thématiques par des méthodes de clustering [devait] pouvoir s'appliquer à l'ensemble des pages Web, sans restriction de taille, de domaine et sans faire appel à d'autres connaissances extérieures que le Web lui-même.* » (*ibid.* p. 221).

Nous avons fait des propositions pour élaborer un profil utilisateur. Un profil utilisateur pouvait être défini par les centres d'intérêt d'un individu (sports, loisirs, domaine professionnel ou scientifique, etc.) et être décliné sous la forme d'une liste de mots. Nous mentionnions que « […] *pour parvenir à une description significative du domaine d'intérêt des utilisateurs, nous faisons l'hypothèse qu'il faut au moins 20 mots pour initier le processus de construction du profil* » et soulignions également que l'effort qui était demandé à l'utilisateur pour définir ces mots ne serait demandé qu'une seule fois, considérant que ses centres d'intérêts étaient stables. Une alternative intéressante aurait été de recourir à des méthodes d'expansion de

requêtes. La morphologie dérivationnelle aurait été ici très utile, mais, ainsi que nous l'avons indiqué, il n'y avait pas d'outils linguistiques dans le laboratoire.

L'expérimentation réalisée par Guillaume Cleuziou a consisté : a) à recueillir des listes de mots auprès de sujets ; b) à interroger des moteurs de recherche avec ces mots dont on cherchait à apprendre les relations contextuelles ; c) à utiliser les pages web retournées par les moteurs pour évaluer les récurrences contextuelles ; d) puis à construire une matrice de proximité. Dans cette matrice, chaque mot est décrit par sa proximité avec les autres mots, et se traduit par une valeur positive (si la proximité est très forte) ou négative (si elle est éloignée). Ce dernier cas arrive par exemple lorsqu'un mot est homonymique d'un autre ou polysémique et qu'il entre dans des réseaux de cooccurrences différents (*avocat / légume* ou *avocat / profession*).

Guillaume Cleuziou a testé plusieurs mesures sur trois listes de mots (*ibid.,* p.230) :

- une liste de 46 mots anglais dans le domaine de l'apprentissage (mots simples et composés tels : machine learning, clustering) ;
- une liste de 44 mots français reflétant les centres d'intérêt d'un sujet (littérature, musique, tourisme dont 3 intrus destinés à tester le classifieur : caméléon, tracteur et horloge) ;
- une liste de 17 mots français dans le domaine du sport.

Deux moteurs de recherche ont été utilisés, un moteur généraliste pour les profils généralistes (Altavista) et un moteur spécialisé pour le profil à caractère scientifique (ResarchIndex). L'interrogation de ces moteurs a permis de recueillir des pages web, appelées improprement « documents », qui ont ensuite été utilisées pour organiser chaque profil. Nous avons analysé les profils obtenus sur le plan linguistique. Ces derniers étaient organisés en « *réseaux d'association révélant aux utilisateurs les plans d'organisation de l'information sur internet* » (*ibid.* p. 233). Nous avons observé que ces plans d'organisation étaient peu homogènes, laissant tantôt paraître des champs lexicaux (par exemple celui de la *tauromachie* ou de la *culture*) tantôt des relations de synonymie (*copains, amis*), ou des associations lexicales révélatrices de l'actualité du moment, etc. Bref, rien qui ne puisse surprendre lorsqu'on travaille sur des « sacs de

mots » et que l'on considère le web comme un corpus, alors qu'il s'agit plutôt d'un « anti-corpus » (Lerat, 2005).

L'expérimentation de 2002 a été riche d'enseignements et a permis à Guillaume Cleuziou d'améliorer sa méthode de classification de mots. Dans (Cleuziou, 2006), l'auteur a introduit la notion de recouvrement de classes. Cette méthode permet de rendre compte que des mots polysémiques, ou homonymiques, appartiennent à plusieurs classes ou qu'un document comporte plusieurs thématiques, ou relève de plusieurs genres. L'auteur a également cherché à évaluer ces résultats suivant les indices de précision et de rappel ainsi qu'il est d'usage en recherche d'information, alors qu'en 2002, l'évaluation des résultats de la classification de mots en profils était humaine, et donc irrecevable pour la communauté scientifique d'appartenance de l'auteur. En ce qui concerne les applications à la recherche d'information, les profils utilisateurs nous paraissaient adaptés à la veille, un profil utilisateur permettant d'initier le processus de veille (Cleuziou et *al*., 2003). Dans ce cadre, il fallait travailler les données, que ce soit en termes de sélection des sources à l'origine de la constitution des corpus et de leur traitement.

A la faveur d'une collaboration avec l'un de nos rapporteurs de thèse, Christian Fluhr du CEA, Guillaume Cleuziou a pu soumettre ses données à des traitements linguistiques, notamment à une analyse morphologique (Cleuziou et *al*., 2003). Quant aux corpus, ils ont par la suite été plus rigoureusement sélectionnés, d'abord en termes de domaine couvert et en termes de traitements. Dans (Cleuziou et al. 2004), l'algorithme développé par l'auteur a été appliqué à deux corpus, le corpus Reuters-21578 et le corpus 20Newsgroup mis à la disposition des apprentistes pour évaluer leurs algorithmes. L'auteur, en utilisant les corpus tests offerts par sa communauté, pouvait désormais soumettre ses algorithmes à des scores de pertinence. Ainsi, la campagne DEFT (Défi Fouille de Textes[85]), lancée en 2005 dans le cadre de la conférence TALN, a-t-elle facilité la mise à disposition des corpus destinée à faire des tests pour diverses tâches de classification. Par exemple, la campagne de 2008 porte sur la classification automatique de documents en genres *(journalistique versus.*

[85] http://deft.limsi.fr/

encyclopédiques) et domaines différents *(art, économie, littérature, politique internationale, politique nationale, problèmes de sociétés, sciences, sports, télévision),* tâche qui a fait l'objet d'une publication très importante pour nous en 2006 (Poudat et *al*., 2006). Pour en comprendre les enjeux, nous devons nous intéresser préalablement aux méthodes issues de la linguistique de corpus.

2.2.2. Caractérisation du genre textuel et linguistique de corpus

2.2.2.1. Enjeux de la production de corpus annotés

Le recours aux corpus électroniques est une pratique courante et ancienne qui remonte à la fin des années soixante et qui entretient des liens étroits avec la lexicographie informatisée. Dès 1964, à l'initiative de Paul Imbs de l'Institut National de la Langue Française (INaLF) une base de données littéraire dénommée Frantext a été développée, qui a permis de réaliser le *Trésor de la Langue Française* (TLF), aujourd'hui connu dans sa version informatisée sous le nom de *Trésor de la Langue Française Informatisé* (TLFI)[86] (Martin, 1995).

Ce qu'il y avait de nouveau en revanche à la fin des années quatre-vingt-dix, c'est le fait que ces corpus ne soient plus des « suites de mots nus » i.e de « simples chaînes de caractères » mais des corpus enrichis par des annotations. L'on parle alors de « corpus annotés » (Habert et al., 1997 : p. 7sqq). Les auteurs de cet ouvrage intitulé *Les linguistiques de corpus* publié en 1997, indiquent qu'il y a eu un regain d'intérêt pour les corpus enrichis et les outils qui permettent de les interroger et de les annoter. Ces annotations sont produites par des outils de TAL : les analyseurs morphologiques produisent des corpus étiquetés et les analyseurs syntaxiques, des corpus arborés. Les corpus électroniques deviennent alors des ressources disponibles pour évaluer les outils : c'est le cas des corpus Reuters et Newsgoup utilisés par la communauté d'apprentissage pour tester les algorithmes de classification. Les corpus sont également le matériau d'une « linguistique outillée » (Habert, 2005), c'est-à-dire une linguistique qui utilise des outils d'analyse de corpus (TAL, lexicométrie, etc.).

[86] http://www.inalf.fr/_ns/index_tlfi.htm

Cette conception du corpus est héritée de la perspective défendue par John Sinclair, l'un des initiateurs de la linguistique de corpus : *« un corpus est un échantillon du langage représentatif de la langue, un vaste ensemble de mots* »… Benoit Habert (2000) indique que la taille ne peut cependant être la seule qualité de ces données langagières (« *gros, c'est beau* »). En effet, les corpus, aussi gros soient-ils ne définissent pas la langue en tant que telle, mais sont des *« réservoirs d'exemples, une manifestation authentique des possibilités de la langue ou d'un discours »* (Tutin, 2010 : p. 4).

Dans le cas où les corpus sont utilisés comme réservoirs d'exemples, Marie-Paule Péry-Woodley note que *« le gigantisme devenu technologiquement facile* » ne libère pas du problème de la représentativité (Péry-Woodley, 1995). Sur ce point, Benoit Habert (2000) mentionne que les critères d'échantillonnage peuvent privilégier les conditions de réception : telles sont les caractéristiques du British National Corpus (BNC)[87] qui représentent une diversité maximale de situations de communication. Les critères peuvent également privilégier les conditions de production. Dans ce cas, les données recueillies sont propres à un domaine (médical par exemple), un lieu (une organisation). Enfin, les critères peuvent privilégier les types de textes, ce qui revient à regrouper des énoncés qui présentent une similarité sur le plan linguistique.

C'est ainsi que nous en sommes venue à la constitution d'un corpus à annoter pour décrire le genre de l'article de revue scientifique.

2.2.2.2. Caractérisation du genre avec des méthodes issues de la linguistique de corpus

En 2001, après avoir rédigé un mémoire de maîtrise sur les outils d'assistance à la traduction multilingue sur internet[88], Céline Poudat débutait un travail de DEA sur le

[87] http://www.natcorp.ox.ac.uk/

[88] Poudat Céline (2000*), Les outils d'assistance à la traduction multilingue en libre accès sur Internet* , mémoire de Maîtrise en Sciences du Langage, G. Bergounioux et V. Clavier, UFR de Lettres.

genre de l'article scientifique qu'elle a soutenu en 2002[89]. Alors que nous participions à une journée d'études de l'ATALA sur le traitement automatique des langues et la linguistique de corpus[90], la communication de Denise Malrieu et de François Rastier sur les genres et les variations syntaxiques publiée dans un numéro de la revue *T.A.L.* (Malrieu et Rastier, 2001) nous est apparue particulièrement éclairante pour le sujet de Céline Poudat. Les auteurs, cherchant à appréhender linguistiquement les normes sociales qui contraignent la notion de genre, mobilisaient les travaux empiriques de Douglas Biber (1988, 1993) qui cherchait à faire émerger des types de textes grâce à un traitement statistique de textes étiquetés. Cet étiquetage ne reposait pas sur un étiquetage morphosyntaxique systématique, mais sur un ensemble de traits qui étaient privilégiés au détriment d'autres, comme les marqueurs de temps et d'aspect, le lieu, les pronoms, etc. Pour Habert et al., il s'agissait d'un « *étiquetage partiel et partial* » (Habert et *al*., 1997 : p. 29). Les statistiques multidimensionnelles étaient ensuite utilisées pour faire apparaître des associations de traits linguistiques qui permettaient de faire ressortir des dimensions caractéristiques de « *types de textes* » : « *Cette démarche permet[tait] la construction inductive d'une typologie de textes basées sur les corrélations effectives entre traits linguistiques* » (*ibid.* p. 30.).

Bien que l'objectif de Denise Malrieu et François Rastier n'ait pas été la reconnaissance automatique de genres, l'étude y contribuait. Pour les auteurs, il s'agissait d'abord d'envisager si les genres prédéfinis se différenciaient suivant un ensemble de variables morphosyntaxiques. Ensuite, il s'agissait d'apprécier le potentiel classificatoire d'un sous-ensemble de ces variables (les dimensions de l'analyse en composantes principales), et enfin, d'explorer l'utilisation de ces sous-ensembles pour différencier des familles de textes à l'intérieur d'un genre. Après avoir distingué différents paliers de caractérisation de la textualité – les discours juridiques versus littéraires–, les champs génériques – à l'intérieur du discours littéraire, le théâtre, la poésie, les genres narratifs –, les genres proprement dits comme la comédie, le roman, les contes, les sous-genres – le roman épistolaire– et

[89] Poudat Céline (2002), *Etude contrastive de l'article scientifique dans une perspective d'analyse des genres,* mémoire de DEA en Sciences du Langage, co-direction G. Bergounioux et V.Clavier, UFR des Lettres, Orléans.

[90] Daille Béatrice et Romary Laurent, *Traitement automatique des langues et linguistique de corpus*, *TAL,* 2 (42), 2001.

enfin les textes de même genres et d'un même auteur. Denise Malrieu et François Rastier expérimentaient leur méthode sur un corpus. Ce corpus était 300 fois plus étendu que celui de Biber, et comportait 2567 textes intégraux (et non des extraits de textes), 164 millions de mots et 250 variables qui se répartissaient dans quatre discours inégalement représentés : scientifique, juridique, essayiste, littéraire. Le corpus avait été préalablement étiqueté avec l'analyseur morphosyntaxique Cordial de la société Synapse-Développement. Les auteurs montraient *in fine* que les variations morphosyntaxiques selon les genres étaient notables, ce qui tendait à montrer que ce niveau de caractérisation qui se situait au palier de la phrase était un compromis acceptable pour caractériser globalement un texte sans tenir compte des paliers supérieurs comme le niveau du discours. Ce résultat était de toute première importance.

Le mémoire de DEA de Céline Poudat s'inscrit dans les traces de ces premiers résultats. Le mémoire identifie notamment l'ensemble des traitements que doivent subir les corpus avant d'être annotés, chaque étape (nettoyage, segmentation, étiquetage, etc.) nécessitant de faire des choix en termes d'outils et de méthodes (symboliques ou numériques). L'auteur dresse un état des lieux des outils de TAL existants (notamment les étiqueteurs probabilistes), des outils statistiques disponibles, et enfin les outils de normalisation et de balisage de corpus comme la Text Encoding Initiative[91] qui permet d'échanger des corpus et de les exploiter grâce à l'adoption d'un système d'encodage (la syntaxe XML) et d'annotation sémantique (la TEI).

A l'issue de la soutenance du mémoire de DEA de Céline Poudat, nous avons contribué à définir son sujet de thèse de doctorat. L'objectif était d'évaluer la « *calculabilité d'une typologie de textes en tenant compte des différents niveaux d'interaction existant entre les types de discours (au sens où ils s'inscrivent dans des pratiques sociales), les types de genres (définis par un ensemble de normes sociales et linguistiques) et les types de textes (rôle des marques d'énonciation)* ». Sur le plan théorique, il s'agissait de réhabiliter le niveau du genre textuel dans une optique contrastive, c'est-à-dire, « *une*

[91] http://www.tei-c.org/index.xml

perspective [qui consiste à créer] des contrastes en explorant la hiérarchie des variables typologiques possibles (e.g. auteur, domaine, genre). On opère donc par cycles de validation, en s'écartant du corpus initial par le biais d'hypothèses contrastives pour finalement y revenir, et le saisir de manière plus pertinente. En ce sens, l'approche adoptée est bien distincte du profilage, qui s'intéresse aux affinités plus qu'aux contrastes, même s'il constate des différences.» (Poudat, 2006 : p. 23). Ainsi que le mentionnait l'auteur, l'une des difficultés pour appréhender les genres sociaux tenait à l'instabilité des critères, les normes variant selon les époques, les cultures et les langues (ce qu'avait montré le mémoire de DEA de l'auteur). La conception du genre héritée de Bakhtine qui avait relié les genres aux pratiques sociales conduisait à appréhender le genre dans ses deux faces, sociale et linguistique :

> *« La notion de genre est une notion biface qui fait correspondre une face interne (les fonctionnements linguistiques) avec une face externe (les pratiques socialement signifiantes). Mais les usagers de la langue utilisent également des termes de classification lorsqu'il y a non coïncidence entre les deux dimensions : tantôt ils privilégient la situation de communication, tantôt les marques formelles. L'instabilité de la relation entre formes et comportements sociaux institutionnalisés est une difficulté centrale pour toute définition a priori des genres. » (Branca-Rosoff, 1999 : p. 116)*

Céline Poudat mentionnait que les démarches les plus répandues consistaient à corréler les phénomènes linguistiques et sociologiques : à partir d'une situation de communication ou d'hypothèses sociologiques spécifiques étaient associées des dispositifs linguistiques particuliers. L'auteur signalait que « *la sélection des descripteurs linguistiques [était] donc orientée par leur interprétation/fonction sociale ultérieure* » (*ibid.* p. 30), ce qui expliquait pourquoi certains marqueurs étaient privilégiés : les marqueurs discursifs (pronoms personnels, modalisateurs) et lexicaux qui sont plus « *facilement interprétables* ». D'après Céline Poudat, cette approche était critiquable puisque était alors exclu un grand nombre de catégories linguistiques pourtant constitutives de la langue et des discours, ce qui était discutable d'un point de vue linguistique. Nous souscrivions totalement à ce point de vue.

L'un des axes forts de la thèse de l'auteur, ainsi qu'elle l'annonçait dans l'introduction de son mémoire résidait dans l'objectif de mettre en œuvre « *un observatoire des genres* » :

> *« La [présente] thèse propose une réflexion méthodologique sur l'élaboration et la mise en œuvre d'un observatoire de genre : après avoir collecté et construit un corpus de 224 articles de revues linguistiques parus autour de 2000 – puisqu'un genre s'observe d'abord en synchronie -, nous avons mis en place une chaîne de traitement exploitant les outils et les méthodes du Traitement Automatique des Langues (TAL), des statistiques textuelles et de la linguistique de corpus en général. » (Poudat, 2006, p. 23)*

Cette thèse a été co-encadrée par Gabriel Bergounioux et François Rastier et a été soutenue en juin 2006[92]. François Rastier considérait avec d'autres auteurs (comme Coseriu qu'il cite[93]) que l'approche en corpus permet d'appréhender les usages linguistiques, l'espace des normes qui les structurent et l'étude de la langue à proprement parler :

> *« Si l'on convient des insuffisances d'une linguistique fondée sur des exemples, pour progresser dans l'étude des genres et de la nature des normes linguistiques qui les structurent, il faut unifier l'étude de la langue et l'étude de la « parole » (au sens saussurien du terme), en étudiant des usages par une linguistique de corpus. » (Malrieu et Rastier, 2001)*

Notre contribution dans cette recherche a résidé dans la mise en œuvre de la chaîne de traitement, notamment lorsque l'auteur a établi un système de catégories morphosyntaxiques propres au genre scientifique pour étiqueter le corpus. Nous avons annoté un corpus suivant les mêmes méthodes de l'auteur dans le domaine de la mécanique, et lui avons cédé notre corpus qui est traité dans le chapitre 8 de la thèse et qui lui a permis de contraster le genre de l'article de revue à deux domaines, la linguistique et la mécanique. Une seule communication a été réalisée en 2006 à

[92] Poudat Céline, *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, thèse de doctorat en linguistique, sous la direction de Gabriel Bergounioux, Université d'Orléans, 2006. [en ligne] http://refef.crifpe.ca/document/these/POUDAT.pdf consulté en mai 2013

[93] « *Le "chaînon manquant" entre la langue et la parole est constitué par l'espace des normes (cf. Coseriu, 1969). Or, seule la linguistique de corpus peut offrir les moyens théorique et technique d'étudier l'espace des normes et de transformer en dualité l'antinomie entre compétence et performance. Pour cela il faut mener une étude comparative, tant des discours que des champs génériques et des genres, voire des styles – c'est là un aboutissement de la problématique de la linguistique comparée.* » (Rastier, 2004a)

l'ATALA sur ce travail en lien avec la recherche d'information[94], la valorisation étant quasiment impossible puisque le cadre théorique était la propriété de l'auteur. Un corpus (même annoté) sans ancrage théorique ne peut être valorisé. Malgré cette déception, nous avons cependant acquis des compétences réelles en matière de constitution de corpus, de traitement et d'annotation qu'il aurait été impossible d'acquérir seule sans émulation tant le travail de balisage et d'annotation de corpus est long, fastidieux et ingrat[95]…

2.2.2.3. Annotation d'un corpus de mécanique

L'approche en termes de contrastes défendue par Céline Poudat dans sa thèse, nous avait conduite à réfléchir aux domaines, notion qui est beaucoup plus familière à la documentation que celle de genre textuel, les genres documentaires ne recouvrant pas les genres textuels[96]. En documentation, les domaines sont appréhendés à travers la terminologie qui est uniquement conçue sous l'angle de l'organisation paradigmatique des langages et de leur coïncidence avec un concept. La terminologie d'un domaine est une affaire de lexique et non de texte, ce qui est fort dommageable à une recherche d'information dans des textes longs, la terminologie variant au fil du texte, et, pour un même auteur, au fil des textes et du temps.

En linguistique, les domaines ont également été considérés comme des *langues de spécialités* ou des *sous-langages* (Harris, 1968). La notion harrissienne de sous-langage s'inscrit dans un système linguistique restreint – langages techniques et scientifiques, métalangages – et permet de formuler des restrictions de sélection d'unités linguistiques particulières – mots, schémas de phrases – ou des marqueurs spécifiques. D'autres travaux plus récents relient les sous-langages à la fois au domaine et au genre (Péry-Woodley et Reyberolle, 1998) : cette perspective permet

[94] Clavier Viviane (2006), « Le genre comme point d'accès au document : analyse comparée de textes scientifiques en mécanique et linguistique », *Journée de l'ATALA Typologies de textes pour le traitement automatique*, 9 décembre 2006, 5p. [En ligne] http://www.atala.org/article.php3?id_article=312

[95] Il s'est passé deux ans environ post-édition comprise pour traiter ce corpus « à temps perdu » entre 2002 et 2004.

[96] Une revue spécialisée par exemple comporte plusieurs genres textuels (la présentation de la revue, les articles, la note de lecture, etc.)

d'identifier des unités fonctionnelles présentant un intérêt pour la recherche d'information, comme la recherche de définitions, conduisant à repérer des faisceaux de traits lexicaux, syntaxiques et typographiques stables (Reyberolle et Péry-Woodley, 1998).

Le domaine que nous avons choisi de décrire était la mécanique : *a priori*, un choix difficile à justifier puisqu'il n'y avait rien en commun avec le domaine de description retenu par Céline Poudat, la linguistique. Nous pensions qu'il fallait bien commencer par un domaine[97], et, comme nous travaillions avec une communauté de chercheurs en mécanique à l'IUT, nous avons pu trouver des corpus. L'un des problèmes que nous avons rencontré est que les sciences de l'ingénieur, comme les sciences de la nature d'ailleurs, ont une activité de publication quasi exclusivement en anglais. Par conséquent, les publications en français sont rares, limitées aux actes de congrès ou de colloques se déroulant en France. Parfois d'ailleurs, même en France, les communications se font aussi en anglais, et l'on peut supposer que les normes de publications sont moins sévères que dans les revues internationales.

Nous avons reçu de l'un de nos collègues de l'IUT, Jean-Noël Blanchard, chercheur à l'Institut de Combustion Aérothermique Réactivité et Environnement du CNRS, un CD-Rom des actes de 2001 du XV^{e} congrès français de mécanique (CFM) de l'association française de mécanique. Nous avons sélectionné un corpus de 49 textes dénommé « corpus CFM », répartis en sept domaines sur les 26 représentés dans les actes du congrès :

- Acoustique (3 textes) ;
- Aérodynamique, Hydrodynamique (6 textes) ;
- Biomécanique (8 textes) ;
- Conception Production (8 textes) ;
- Dynamique des structures et des machines (7 textes) ;
- Ecoulements polyphasiques (8 textes) ;
- Endommagement – rupture – fatigue (4 textes) ;

[97] Cet objectif s'est d'ailleurs réalisé puisqu'en 2007, nous avons participé à la constitution d'un corpus de données scientifique et technique qui rassemblait plusieurs genres et plusieurs domaines (cf. SCIENTEXT, voir en deuxième partie de ce mémoire).

- Environnement, Milieux poreux (5 textes).

Tous les traitements ont fait l'objet de procédures[98] rédigées par Céline Poudat que nous testions et adaptions aux articles de mécanique et qui ont fait l'objet de l'annexe 5 de la thèse de l'auteur. Nous avons suivi cette chaîne de traitements, de la numérisation des articles scientifiques à leur encodage. La norme d'encodage pour décrire ces sources a été la Text Encoding Initiative (TEI), pour laquelle il était nécessaire de définir les éléments pertinents propres à décrire les genres textuels, de choisir les balises proposées par la TEI et de définir la granularité de la description. Nous voulions en effet produire des métadonnées de différents niveaux : des informations qui décrivent les mentions bibliographiques – nom des auteurs, titre, affiliation –, la structure des articles scientifiques – titres, résumés, mots-clés, notes –, la diversité des sémiotiques – les figures, schémas, textes – l'intrication des langages – langage mathématique et langage naturel –, la diversité des catégories – morphosyntaxiques, énonciatives, sémantiques. La première étape a consisté à numériser les articles scientifiques fournis au format pdf. Nous avons utilisé les moyens du bord (logiciel d'OCR, Omnipage) et avons corrigé les erreurs en utilisant des expressions régulières (Procédure nettoyage de corpus). Toutes les figures – schémas, graphiques, tableaux, etc. – ont été consignées dans des fichiers séparés en vue de leur balisage ultérieur. Il s'est avéré que toutes les équations et formules mathématiques avaient disparu après l'océrisation, ce qui nécessitait de rééditer les équations : nous l'avons fait pour une partie du corpus, ensuite nous avons cessé ce travail titanesque pour simplement introduire une balise TEI « formula » sans mentionner d'attributs.

La deuxième étape a consisté à baliser les textes suivant les conventions de la TEI, (Procédure balisage de corpus). Nous avons travaillé sur des fichiers word, afin de baliser la structure en sélectionnant les niveaux de description des documents. Nous avons balisé les titres, les résumés, les mots-clés, les abstracts, les key-words, les différents niveaux de structure du document, y compris les notes de bas de page et les références bibliographiques. En outre, nous nous sommes attachées à baliser les

[98] Le site *Texto ! Textes et Cultures* présente un ensemble de procédures et de méthodes de traitements de corpus dans la rubrique « Corpus et méthodes » http://www.revue-texto.net/index.php?id=62.

figures et les langages mathématiques qui ponctuaient le texte (équations, valeurs numériques, constantes, etc.) et qui pouvaient apparaître soit de manière détachée avec un retrait dans le texte, soit directement au sein de la phrase, ce qui, comme on peut se l'imaginer était un réel problème pour la reconnaissance automatique. Ce balisage était parfois complexe, puisqu'il nécessitait de nommer les types de figures. La caractérisation des formules mathématiques exigeait une connaissance du langage mathématique, ce qui n'était pas notre cas.

La troisième étape a consisté à transformer ces fichiers ainsi balisés au format .txt, puis au format .xml en validant la conformité du document XML avec Cooktop, un éditeur XML gratuit. A titre d'exemple, nous présentons en annexe un document balisé suivant ces normes (Annexe 1).

La quatrième étape a consisté à étiqueter les textes. Ainsi que nous l'avons indiqué précédemment, l'approche robuste devait fournir une solution et une seule pour chaque forme du texte[99], ce qui avait conduit Céline Poudat à choisir des méthodes probabilistes pour étiqueter les textes (c'est pourquoi l'on parle de tagger ou d'étiqueteur et non d'analyseur). Le recours à ces méthodes s'était imposé après que l'auteur eut testé le logiciel commercial Cordial, qui, bien que proposant un grand nombre de variables (plus de 200) s'était révélé peu adapté à la description du discours scientifique, et en particulier aux articles de revue en linguistique. L'auteur indiquait dans sa thèse que « *l'étiqueteur propose [ainsi] de nombreuses catégories sémantiques ambiguës et peu transparentes (e.g. noms 'abstraits' / 'concrets', 'humanoïdes', etc.), et considère les paragraphes précédés de tirets comme indices de 'dialogue', ce qui illustre l'orientation du logiciel vers la description des textes littéraires.* » (*ibid*., p. 85) C'est pourquoi l'auteur avait développé son propre jeu d'étiquettes, puisqu'il n'existait aucun outil d'étiquetage automatique spécifiquement adapté à l'observation du discours scientifique.

La mise au point d'un jeu d'étiquettes dédiées au discours scientifique s'était révélée délicate. En effet, d'un côté, il était nécessaire de concevoir des outils généralistes

[99] Céline Poudat note que *robustesse*, *efficacité*, *précision*, *adaptabilité* et *ré-emploi* sont les qualités que doit présenter un étiqueteur d'après (Cutting et al., 1992).

réutilisables pour la classification de données textuelles relevant du genre de l'article de revue. Pour cela, un jeu d'étiquettes restreint aurait été suffisant (50 étiquettes). De l'autre côté, moins il y a d'étiquettes, plus le pouvoir de discrimination des étiquettes sur un corpus aussi homogène que celui qu'elle avait constitué en linguistique était faible. C'est pourquoi, l'auteur avait fait le choix de « *variables plus fines, plus adaptées aux caractéristiques des textes et plus interprétables* ». L'auteur a proposé 15 classes morphosyntaxiques conduisant à 145 variables dédiées au discours scientifique. Ce jeu d'étiquettes a été appliqué à un corpus d'entraînement en linguistique avec plusieurs étiqueteurs (Brill Tagger, MBT Tagger, Tree Tagger et TnT Tagger) et, après correction manuelle des sorties des étiqueteurs, c'est finalement TnT Tagger qui a été retenu par l'auteur.

Notre corpus a également été soumis à un étiquetage, mais comme la mécanique n'avait pas fait partie du corpus d'entraînement, j'ai dû corriger manuellement les sorties des 49 textes, ce qui a représenté, un certain nombre de semaines de travail. L'on pourra se reporter à l'annexe 2 pour voir un extrait du corpus étiqueté. Par la suite, les catégories étiquetées ont été intégrées à la structure du document XML grâce à l'aide de Sylvain Loiseau, à l'époque doctorant dans le laboratoire de François Rastier et qui était intervenu aux séminaires de formation à la TEI à Orléans. Le corpus de mécanique a ensuite été contrasté au corpus de linguistiquel (cf. chapitre 8 de la thèse de Céline Poudat).

Concernant les variables, Céline Poudat note que 4/5ème des variables sont communes aux deux domaines, ce qui confirme la pertinence des catégories pour identifier la dimension générique des textes : l'on peut alors en déduire que les domaines sont appréhendés par le lexique et le genre par les catégories morphosyntaxiques. La perspective adoptée étant différentielle, cette méthode a également permis d'identifier les variables absentes (*ibid*. p. 308sqq). L'auteur constate qu'un cinquième des étiquettes présentes en linguistique est absent du corpus de mécanique : si certaines absences sont liées au domaine traité comme les marqueurs grammaticaux propres à la linguistique, le *je* est totalement absent (c'est le *nous* qui est utilisé en mécanique) ; certains temps grammaticaux sont inemployés (le passé simple et certains temps pour les verbes de modalité), toutes les

ponctuations ne sont pas utilisées (pas de points virgules, ni de points de suspension) ; aucun connecteur de concession n'a été relevé, ni d'adjectif réflexif (*même,* comme dans *lui-même).* L'auteur en tire des conclusions sur les pratiques d'écriture scientifique observées en linguistique et mécanique. Par ailleurs, en analysant les deux premiers facteurs d'une analyse en composantes principales (ACP) réalisées sur le corpus de mécanique (*ibid.* p. 310) comparés à la même analyse réalisée sur le corpus de linguistique (ibid. p.), Céline Poudat constate que les organisations morphosyntaxiques des variables sont proches, mettant en évidence deux pôles *formalisation* versus *mode de narration historico-narratif.* Elle montre également que les marqueurs énonciatifs de la rhétorique scientifique (temps du présent, impératif, on, les modaux du conditionnel, les connecteurs de la spatialité et de la présupposition) sont fortement corrélés aux indices de la formalisation. Elle mentionne également des différences sur le rôle des pronoms personnels, le *on* étant souvent utilisé en mécanique, le *nous* en linguistique.

Concernant la comparaison des structures génériques des textes de linguistique et de mécanique, l'auteur mentionne que les articles de linguistique ne sont pas soumis à une structure rhétorique très normée. Inversement, les articles de mécanique sont conformes à la structure dénommée IMRaD (Information, Methods, and Results and Discussion), caractéristique des disciplines empiriques. Ce constat a conduit l'auteur à décrire les introductions et les conclusions des articles linguistiques afin d'en cerner les spécificités.

2.2.3. Applications à la recherche d'information

2.2.3.1. Le genre comme « plan d'organisation de la textualité » et « point d'accès au document »

En 2006, nous avons présenté une communication à l'ATALA dans laquelle nous voulions montrer que *« la recherche d'information aurait intérêt à prendre en compte d'autres paliers de description que la seule thématique pour affiner les critères de recherche.* » (Clavier, 2006). Nous mettions en avant le genre comme « *plan*

[important] *d'organisation de la textualité* » et comme « un *point d'entrée* [possible] *au document* » (Clavier, 2006).

Nous avions situé notre propos dans le cadre de la recherche d'information spécialisée qui constituait « *un terrain d'observation privilégiée des usages liés à la documentation scientifique* ». Nous faisions alors le constat suivant : « *C'est […] l'étude de la terminologie qui a été, et est encore, l'une des approches les plus couramment adoptées pour aborder les discours scientifiques. Ces travaux ont fait l'objet de nombreuses applications en recherche d'information. Le recueil de la terminologie dans les textes spécialisés est alors le point de départ d'une représentation des connaissances sous forme de concepts et de relations au sein de thesaurus et d'index.* » Or, nous indiquions qu'une représentation qui s'appuie uniquement sur une organisation décontextualisée des lexiques (thématique, terminologie) était préjudiciable à l'interprétation des termes, laquelle différait suivant leur lieu d'apparition dans la structure du document :

> *« Il apparaît pourtant que le relevé de termes que l'on érige en listes ou nomenclatures indépendamment de tout contexte discursif présente de nombreuses limites. (Clavier, 2006 : p. 3)*

Nous illustrions ce fait par des exemples issus du corpus de mécanique et concluions : « *L'on observe ainsi que le sens d'un même mot varie au fil du texte. Apparaissant en début de texte, un mot pourra être problématisé, alors qu'à la fin, il sera défini. Il est donc important de considérer que le genre a aussi un impact sur la caractérisation du lexique de spécialité.* » (*ibid.)* Ce point de vue nous conduisait à nous inscrire dans le sillage de François Rastier (1989) qui choisissait le genre comme dimension du texte à décrire, plutôt que de se placer au niveau du lexique ou de la phraséologie, afin de rendre compte « *des contraintes globales sur le local* ». Nous indiquions que « *les textes écrits utilisent de façon normée le système graphique, la langue, la structure schématique, l'organisation linéaire des textes.* » et soulignions le rôle du genre pour la recherche d'information, dont on peut signaler ici que, pour la première fois dans nos propos, l'expression faisait référence à une activité sociale *« Ces contraintes témoignent d'une codification socio-discursive et culturelle et, il est bien établi que les genres textuels sont révélateurs de ce système de normes (Malrieu et Rastier, 2001). Si la*

connaissance des genres est indissociable des activités d'écriture, elle conditionne aussi la réception et l'interprétation des textes. De ce point de vue, le genre présente un intérêt pour la recherche d'information. »

Les méthodes que nous avons utilisées avec Céline Poudat s'inscrivent dans le sillage des travaux de Benoît Habert et al. sur le profilage de textes[100] (2000) et de ceux de Douglas Biber, dont l'ouvrage précurseur publié en 1988 *Variation assross Speech and Writing* posait les principes d'une méthode inductive destinée à faire émerger des typologies textuelles à partir de critères linguistiques. Nous recourions ainsi à la linguistique de corpus, aux langages (XML), à la norme TEI pour baliser la structure d'un document et son contenu. Ces méthodes étaient susceptibles de (ré-)concilier les perspectives documentaire et sémiotique grâce à la prise en compte du genre comme lieu de régulation sociale pour décrire les textes. Cependant, notre étude manquait d'ancrage social : Quels usages étaient faits du genre en recherche d'information ? Dans notre communication, nous y faisions référence, puisque nous indiquions : « *l'observation des pratiques des chercheurs montre que ces derniers attachent tout autant d'importance à la pertinence thématique d'une information qu'au type de genre textuel auquel il se rattache* ». Mais cette assertion était illustrée par des exemples tirés d'une expérience intuitive non étayée scientifiquement : « *Par exemple, un chercheur ou enseignant-chercheur, va, dans le cadre de son activité, sélectionner des documents qui seront des articles de recherche, des résumés, des procédures, des analyses critiques d'articles, des rapports de recherche, des cours en lignes etc. Ces discours relèvent tous d'un genre particulier et l'usage de ces genres semble même relativement codifié*. » Cette justification par les usages sociaux « supposés », guidés par les représentations que se font les concepteurs des usagers des dispositifs qu'ils développent, était caractéristique des approches techniques que nous avions toujours connues. Mais à cette époque, nous n'étions pas encore armée méthodologiquement et théoriquement pour mener des études d'usage bien que nous commencions à en percevoir l'intérêt puisque nous venions d'obtenir une mutation à Grenoble en 2004.

[100] « *Nous appelons profilage de textes un bilan quantitatif fondé sur des indices linguistiques d'emploi du vocabulaire, mais aussi de catégories (morphosyntaxiques, syntaxiques, sémantiques, structurelles) et de patrons morphosyntaxiques, etc., dans les parties d'un corpus, pour regrouper ensuite ces parties en sous-ensembles homogènes sur ces points. Ce bilan doit également permettre de positionner un nouveau texte par rapport aux regroupements déjà obtenus*. » (Habert et al. 2000)

2.2.3.2. Classification supervisée de documents en genre et domaine

Disposant de corpus spécialisés annotés dans des domaines et des genres différents, nous étions alors en mesure d'évaluer les méthodes d'apprentissage pour réaliser des typologies textuelles. C'est ainsi qu'un article écrit en collaboration avec Guillaume Cleuziou, Céline Poudat et nous-même a été publié dans la revue *Document numérique* en 2006 (Poudat et *al*., 2006). Considérant que les domaines pouvaient être appréhendés au niveau du contenu lexical et les genres au niveau morphosyntaxique, l'article se proposait de « *mesurer l'impact et la complémentarité de deux niveaux de description des textes pour la classification* » (*ibid*., p. 61) L'étude visait à évaluer le caractère discriminant des descripteurs morphosyntaxiques et lexicaux pour classer les textes en genres et domaines à partir d'un corpus « pilote » de taille restreinte.

Nous avions situé l'étude dans le cadre de la classification supervisée, également appelée catégorisation. Les méthodes de catégorisation permettent de prédire la catégorie d'un objet à classer par apprentissage. Pour classer des textes, les méthodes les plus courantes sont les suivantes (cf. Ibewke-SanJuan, 2007 : p. 94sqq) : la méthode des k-plus proches voisins utilisée en recherche d'information pour fournir à l'utilisateur des documents similaires à ceux qui l'intéressent ; la méthode des machines à vecteurs supports (SVM) s'appuie sur des représentations vectorielles (présence / absence) et est utilisée pour effectuer des classements binaires de textes (par exemple des classements d'opinion) ; les classifieurs bayésiens naïfs sont utilisés pour prédire la valeur d'un objet sur des principes probabilistes ; les arbres de décision permettent une catégorisation binaire des textes tout en acceptant de prendre des valeurs catégorielles (contrairement aux méthodes SVM par exemple qui ne tolèrent que des variables numériques) (*ibid.* p. 106). Quelle que soit la méthode utilisée, « *le problème se ramène à celui de la détermination des caractéristiques les plus importantes des textes, usuellement à travers les mots qu'ils contiennent* » (Ibekwe-SanJuan, 2007 : p. 94).

Les expérimentations proposées dans (Poudat et *al*. 2006)[101] avaient pour objectif d'évaluer l'influence de chaque type de description sur la classification (précision du

[101] Méthode également exposée dans le chapitre 8 de la thèse (Poudat, 2006 : p. 312sqq)

classifieur). En outre elles permettaient d'observer comment les deux ensembles d'attributs (lexical et morphosyntaxique) se combinaient dans un même classifieur. Guillaume Cleuziou a utilisé deux méthodes : la classification par SVM et par arbres de décision. La classification de textes en genre n'était pas une nouveauté et différents types de marqueurs caractéristiques du genre avaient déjà été relevés dans la littérature. Généralement, les descripteurs retenus se résumaient en une combinaison de traits structuraux : catégories grammaticales, ponctuation, longueur des phrases, complexité syntaxique, *etc*. (Kessler et *al*., 1997), (Lee et Myaeng, 2002). Nous indiquions que « *La disponibilité de documents numériques sur internet [avait] récemment contribué à remettre le genre sur le devant de la scène (Prime-Claverie et al., 2002). C'est ici la portée typologisante du genre qui est « utilisée » à des fins classificatoires, d'autres travaux plus anciens en avaient déjà posé les principes (Karlgren et Cutting, 1994).* ».

La méthode décrite dans l'article précisait les choix de descripteurs. Ont été retenues comme variables caractéristiques du domaine, les substantifs qui supportent la terminologie, et, comme variables caractéristiques du genre scientifique, une sélection de 136 variables comme les indices de structuration, les connecteurs, les symboles, les verbes de modalités, etc. Le corpus pilote contenait 371 textes scientifiques français dans deux domaines (322 en linguistique, 49 en mécanique), trois genres (273 articles, 45 présentations de revues et 53 comptes rendus), tous les textes étant étiquetés avec le logiciel TnT tagger. Afin d'évaluer les tâches de classification, deux sous-corpus avaient été établis :

- l'un qui se composait uniquement des articles, mêlant les deux domaines (dénommé ART-corpus) : la tâche de classification devait permettre de distinguer les domaines ;
- et l'autre qui ne comportait que des textes de linguistique (dénommé LIN-corpus) la tâche de classification devait permettre de distinguer les genres d'un même domaine.

Concernant la première tâche, nous faisions le constat suivant: *« Les résultats obtenus avec la méthode SVM montrent clairement et contre toute attente, que les variables*

morphosyntaxiques sont plus discriminantes que les variables lexicales. De plus, on note qu'une utilisation conjointe des deux types de variables est globalement plus efficace que chacun des deux ensembles choisi séparément. » (*ibid.* p. 68). Concernant la seconde tâche, nous constations que *« Les résultats obtenus avec le classifieur SVM confirmeraient l'hypothèse selon laquelle les genres sont effectivement corrélés au niveau morphosyntaxique: le taux de précision obtenu est plus élevé avec les jeux de variables comprenant des attributs morphosyntaxiques qu'avec les variables lexicales uniquement. »* (*ibid.* p. 69).

La seconde question qui était posée dans cet article résidait dans l'identification des variables discriminantes qui intervenaient dans les tâches de classification. Pour cela, les arbres de décision avaient été utilisés car ils permettent de « visualiser » la tâche de classification, les règles de classification apparaissant sous la forme d'un arbre dont les feuilles indiquent l'appartenance à une classe.

Pour classer les textes en domaines, les résultats montraient que les variables lexicales discriminantes appartenaient toutes au domaine scientifique de la mécanique, les textes de linguistique étant différenciés de manière négative. Cette discrimination par des termes de mécanique s'expliquait d'abord par la taille plus importante des textes de linguistique qui augmentait le nombre et la diversité des descripteurs ; ensuite, par les textes de mécanique qui semblaient plus homogènes au niveau lexical. (*ibid*., p. 71)

En revanche, pour les descripteurs morphosyntaxiques, c'est l'inverse qui se produisait, puisque c'étaient les variables linguistiques, et en particulier les prépositions, les pronoms personnels les marques de renvoi, qui permettaient de discriminer les textes de linguistique de la mécanique. Enfin, concernant les classifications mixtes, l'étude montrait que les variables lexicales apparaissaient en premier, mais c'étaient pourtant les variables morphosyntaxiques qui permettaient d'affiner la classification – et, c'était là l'intérêt majeur de cette étude –, et d'améliorer les typologies textuelles en domaines.

Pour classer les textes en genre, l'étude montrait le rôle prépondérant des variables lexicales qui permettaient de discriminer la quasi-totalité des comptes rendus et des présentations de revues. L'étude indiquait alors que « *les articles [étaient] donc classés relativement à l'absence de marqueurs caractéristiques des autres genres* » (*ibid*. p. 73), tels que les mots « chapitres », « contributions », etc. Ces résultats étaient confirmés par d'autres auteurs (Lee and Myaeng, 2002), qui stipulaient que certains indices lexicaux étaient propres au genre. L'étude soulignait néanmoins que les éléments lexicaux n'étaient pas aussi efficaces pour distinguer les genres que les domaines. A ce titre, notre étude soulignait le rôle déterminant des marques de structuration textuelle, qui apparaissent en premier dans les arbres de décision et permettaient de discriminer d'un côté les genres de l'article et des présentations de revues, de l'autre les comptes rendus, qui ne sont pas structurés.

2.3. Bilan critique

2.3.1. Les apports à l'organisation des connaissances

Un premier apport réside dans l'introduction d'annotations linguistiques pour améliorer des méthodes mathématiques de classification de textes. Le principal enjeu nous concernant, résidait dans la description des connaissances propres au genre textuel dans le but de mettre en correspondance des textes classés thématiquement avec des requêtes.

Tout comme dans le chapitre précédent, notre contribution porte sur le choix d'un niveau d'organisation et de représentation des connaissances pour l'indexation, étape qui précède la tâche de classification à proprement parler. Cependant, à la différence du premier chapitre, les connaissances ne décrivent pas des unités linguistiques infra-phrastiques (les morphèmes), mais des unités supra-phrastiques (les textes) ; les connaissances ne s'appuient pas uniquement sur la langue pour décrire ces unités (morphosyntaxe, sémantique) mais sur les normes sociales et linguistiques qui contraignent les genres textuels ; en outre, contrairement au premier chapitre, les connaissances recueillies ne sont pas organisées et structurées dans des ressources

(dictionnaires et grammaires) mais sont annotées directement dans les textes qui servent de données empiriques pour tester les algorithmes de classification.

Le cadre expérimental a permis de faire varier plusieurs paramètres :

- l'origine des données : le web et les bases de données spécialisées ;
- le domaine de l'information : l'information grand public et l'information scientifique et technique ;
- la nature des descripteurs : lexicale et morphosyntaxique ;
- les applications : la veille et la recherche d'information ;
- les méthodes de classification : supervisée et non supervisée ;
- la nature des objets textuels à classer : les mots, les genres et les domaines.

La progression de notre travail s'est faite en trois temps. Tout d'abord, nous avons cherché à donner du sens à la notion de « similarité » en linguistique. Jusqu'alors, nous concevions que des énoncés pouvaient être similaires s'ils étaient issus de formes canoniques (sous-jacentes) identiques : c'est ainsi que nous avions appréhendé la synonymie lexicale entre des mots construits ou la paraphrase entre des énoncés grâce à l'analyse automatique de discours, faisant l'hypothèse que les morphèmes portaient les significations. Dans les modèles mathématiques issus de Gérard Salton, les textes sont transformés en des représentations vectorisées, et les mesures considérées sont d'une tout autre nature : elles sont statistiques et non pas linguistiques. Suivant cette conception, il n'y a plus d'ordre dans le texte qui est réduit à un « sac de mots ». La pertinence est évaluée en tenant compte des positions respectives d'un document par rapport à une requête et est estimée par une distance, c'est-à-dire au sens mathématique, une mesure définie dans un espace euclidien, qui traduit une proximité sémantique. La prise en considération des associations de termes – cf. la notion de co-occurrence – permet de réduire les cas d'ambiguïtés et améliore la pertinence. D'une manière générale, nous avons pris conscience des apports considérables qu'apportent les statistiques à l'analyse des textes et à leur comparaison.

Ensuite, nous avons intégré les corpus dans nos méthodologies, avons appris à les constituer et à les traiter linguistiquement. L'approche qui a présidé au recueil d'un

corpus de linguistique (par Céline Poudat) et d'un corpus de mécanique (par nos soins) est liée à des objectifs de recherche visant à éclairer une étude linguistique sur les genres. Ce corpus, ainsi que le signale Céline Poudat reprenant les termes de Sylvain Auroux (1998) est un « observatoire » des genres qui se révèle alors un lieu d'expérimentation et de validation d'hypothèses en linguistique. C'est aussi un objet documentaire au sens où il a permis la production de métadonnées descriptives et bibliographiques : en ce sens, le recours à un codage XML et à la norme TEI a fait de ce corpus, une archive documentaire.

Cette approche en corpus nous a permis de faire le lien entre le texte et le document, le texte étant considéré comme un niveau de description adéquat pour cerner la signification. Jusqu'à cette époque en effet, notre héritage structuraliste nous avait conduite à privilégier les unités de description internes à la phrase, dans une optique logico-formelle et à travailler hors discours. Avec le texte, le corpus est conçu comme un objet d'utilité sociale et scientifique :

> *« Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. » (Rastier, 2004a)*

Au moment où nous publions les résultats de ce travail sur corpus (Clavier, 2006) et où Céline Poudat soutenait sa thèse (Poudat, 2006) paraissait un ouvrage dont nous n'avions pas eu connaissance sur le moment et qui devait connaître un retentissement important dans la communauté des sciences de l'information et de la communication : *Le document à la lumière du numérique* de Roger T. Pédauque, un collectif de 175 chercheurs (Pédauque, 2006). Les auteurs déploraient alors que les théories du texte fussent minoritaires en linguistique[102] alors même que « *l'enjeu actuel [était] de dépasser l'héritage référentialiste pour fonder une sémiotique du texte tirant profit à la fois des possibilités d'analyse de la linguistique de corpus et de la*

[102] « *Le* texte *étant à l'évidence une dimension du langage, on s'attendait à ce que les textes concrets constituent l'objet empirique essentiel de la linguistique. Ce n'est pas le cas. Le texte n'a jamais vraiment constitué une unité minimale d'étude pour la linguistique, en partie parce que celle-ci souhaitait se démarquer de la littérature, jugée non scientifique.* » (Pédauque, 2006 : p. 98)

problématique de l'hypertextualité et des techniques multimedias » (*ibid.* p, 101). Les auteurs donnaient alors les lignes directrices de ce que nous pourrions qualifier de « programme » et dont nous pensons qu'il correspond à la perspective que nous avons adoptée dans la suite des propositions de François Rastier :

> *« Il s'agit de fonder le texte comme unité d'analyse intrinsèque avec ses corrélats méthodologiques d'intra- et d'intertextualité. A la différence de la phrase ou les parties du discours ont des signifiants isolables, le texte ne peut être segmenté en unités sémantiques directement identifiables. L'intratextualité ne se réduit pas évidemment à quelques mots clés qui en subsumeraient la thématique* ***; Le sens d'un texte résulte de connexions de traits sémantiques qui se manifestent à différents paliers du texte (phonème, morphèmes, lexies, prosodie…). Ces traits relatifs à la fois au fond et à la forme sont identifiés par l'interprétation dans la construction du sens, la signification. »*** *(Pédauque, 2006 : p. 101) – C'est nous qui soulignons.*

Les auteurs relèvent alors le rôle fondamental que joue le genre, puisque tout texte appartient à un genre dans une culture donnée :

> *« Le genre codifie la production et l'interprétation du texte, il relève du social et non du fonctionnel de la langue. […] L'analyse des textes implique donc la comparaison de textes de même genre, principe trop peu pratiqué dans la linguistique de corpus ». (ibid., p. 101)*

Enfin, nous avons montré quels étaient les apports de la fouille de textes et de la linguistique de corpus à la recherche d'information. Les méthodes de classification ont été appliquées aux mots pour organiser des profils utilisateurs, ce qui présente un intérêt notamment pour la veille (Clavier et *al*. 2002 ; Cleuziou et *al*. 2003 ; Cleuziou et *al*. 2004a et 2004b). Elles ont également été appliquées aux textes pour les classer en genre et domaine (Poudat et al. 2006). Pour la recherche d'information, le niveau de structuration des documents pour contextualiser la terminologie nous paraissait essentielle à l'interprétation.

Pour terminer, la découverte des méthodes statistiques nous ont vivement intéressée. Nous les avons mises à contribution pour évaluer le niveau d'appropriation des savoirs d'étudiants en situation d'écriture. Auparavant, en 2002, nous avions créé un cours pour les étudiants orléanais de licence puis de maîtrise en sciences du langage sur les statistiques textuelles et leurs applications pour les textes littéraires

(recherche de paternité d'auteurs, stylométrie, analyse de la richesse lexicale, etc.)[103]. En 2004 et 2005, nous avons appliqué ces méthodes dans un contexte pédagogique au département de génie mécanique et productique de l'IUT d'Orléans (Clavier et Guet, 2004) et (Clavier et Lafont-Terranova, 2005). Dans ces études, nous défendions l'idée qu'il est nécessaire d'introduire dans les filières technologiques des cours d'enseignement du vocabulaire de spécialité en lien avec les enseignants d'expression et communication et les enseignants de spécialités, en l'occurrence ici Jean-Michel Guet, professeur en sciences des matériaux au département GMP d'Orléans.

> *« Enseignant l'Expression et Communication (EC) et les Sciences des Matériaux (SDM) à l'IUT d'Orléans, nous rendons compte d'une expérience initiée en 2001 qui a été menée auprès d'étudiants de deuxième année en génie mécanique et productique. Elle vise à inscrire les pratiques de l'écrit dans un domaine de spécialité qui participe à la formation Scientifique et Technique (S&T) des futurs techniciens. Le but est de favoriser la construction de savoirs disciplinaires et leur appropriation par le biais de l'écriture. » (Clavier et Guet, 2004 : p. 95)*

Cette expérience posait de manière centrale la question de la transversalité des apprentissages, cet enseignement ne pouvant se faire sans une implication active des enseignants des spécialités concernées afin *« de réunir les conditions d'appropriation du vocabulaire, de sa contextualisation et de sa conceptualisation »*. (Clavier et Terranova, 2005). C'est dans ce cadre que nous avons utilisé des méthodes statistiques pour comparer le vocabulaire des productions des étudiants (des synthèses de documents) avec celles des experts (des articles de revue) :

> *« L'analyse factorielle des correspondances issue des techniques d'analyse de données multidimensionnelles permet de considérer le vocabulaire intégral de chacun des textes et de mesurer la « connexion » (ou distance) entre les textes et de visualiser les résultats sur une carte. Un mot contribue à rapprocher deux textes s'il est commun aux deux et à augmenter la distance s'il ne se rencontre que dans un seul » (Clavier et Lafont, 2005)*

L'analyse des fréquences révélait globalement une bonne reprise des thèmes principaux des experts, ce qui revenait à dire que les étudiants maniaient à peu près la même proportion de termes spécialisés que les experts. Mais nous observions des différences entre les productions. Certaines révélaient des difficultés rédactionnelles :

[103] Intitulé « Statistiques textuelles » dont un chapitre sur la richesse lexicale et les modèles hypergéométriques utilisés notamment dans le logiciel Hyperbase.

elles étaient marquées par un fort taux de répétivité ; d'autres négligeaient les contraintes de genres : elles étaient marquées par un texte non construit, une sorte de « texte-liste ; enfin, certaines se situaient à un niveau de vulgarisation marqué par un vocabulaire peu spécialisé.

2.3.2. Les limites

Le contexte de réalisation de ces publications était tout à fait différent du premier, et il est difficile de juger des résultats à l'aune des mêmes critères que précédemment. L'environnement institutionnel se révélait à la longue épuisant, tant les cadres théoriques et méthodologiques des deux laboratoires de recherche auxquels nous appartenions étaient différents. De surcroît, nos enseignements en expression et communication n'étaient pas non plus en phase avec notre recherche, si bien que nous avons dû créer un certain nombre de cours pour enseigner en second et troisième cycle. Notre implication pédagogique et administrative à l'IUT était très forte, puisque nous avons été un temps chargée de mission. Nous étions en permanence écartelée entre les représentations que se faisaient les collègues de notre discipline (ou était-ce notre cas particulier ?) : spécialiste des technologies d'information et de communication puisque nous étions chargée de mission aux NTIC ; experte en communication écrite et orale dans le cadre de nos enseignements en expression et communication ; linguiste selon les informaticiens, taliste selon les linguistes… La position que nous avons tenue avec les doctorants, à mi-chemin entre l'encadrement et la collaboration, était assez inconfortable également et s'est traduite par un déficit de publications, alors même que nous avions constitué un travail de corpus très important, et que nous nous étions efforcée de travailler le lien entre la classification de textes et la recherche d'information.

Si l'on se penche à présent sur les limites de ce travail, l'on peut en évoquer deux : la première relève d'une critique généralement opposée au courant de l'*information retrieval* et se résume en l'absence de prise en compte de l'usager et du contexte d'utilisation des systèmes de recherche d'information pour les évaluer. La seconde est liée à la méthodologie de la linguistique de corpus qui se révèle lourde à mettre en œuvre et qui présente également des points faibles.

Concernant le premier point, nous avons déjà relevé dans le premier chapitre de ce document, des travaux critiquant les évaluations des modèles de recherche d'information menées en laboratoire et qui ne sont jamais confrontés aux usages réels (Chaudiron 2004a, 2004b ; Ihadjadene et Chaudiron, 2008). Or, nos travaux étaient susceptibles de recevoir les mêmes critiques. Pour la première fois dans nos publications en effet, nous évoquions le lien entre l'organisation des données textuelles, la lecture, la visualisation, l'interprétation pour évaluer la pertinence d'un système de recherche d'information qui restitue des informations « classées » à un utilisateur (Clavier et al., 2002). On ne peut que regretter qu'à cette époque nous n'ayons pas imaginé mettre en oeuvre des expérimentations ou des entretiens pour valider ces hypothèses… Mais ainsi que nous l'avons signalé *supra*, nous ne connaissions aucune méthodologie issue des sciences sociales et cognitives. Et pour tout dire, nous pensions que cela n'était pas de notre ressort mais de celui des psychologues ergonomes et des sociologues des usages.

Concernant le second point, le travail sur corpus a nécessité un déploiement de techniques extrêmement lourd pour recueillir des textes caractéristiques du genre de l'article (ou tout au moins proches), les numériser, les étiqueter[104], les corriger, les annoter avec la norme TEI. Il est à se demander si de tels traitements sont envisageables à grande échelle, alors même que le web sémantique aurait tout intérêt à intégrer des approches plus contextualisées. D'une manière générale, de nombreuses disciplines ont convergé autour des corpus et de leurs traitements, ainsi que l'évoquent Nathalie Aussenac-Gilles et Anne Condamines à propos de la linguistique de corpus, la terminologie, l'informatique et la recherche d'information (Aussenac-Gilles et Condamines, 2007). En fonction de l'application, plusieurs familles de traitements interviennent comme les méthodes numériques, symboliques, de visualisation de données. La maîtrise de l'ensemble de ces méthodes ne va pas de

[104] Précisons qu'au début de la thèse, il existait des corpus étiquetés ou arborés mais la plupart étaient en anglais (Brown, LOB, Helsinki, Suzanne, etc.). En français, les corpus existants étaient rarement dans le domaine public (cf le corpus Menelas sur les maladies coronariennes, le corpus Mitterand constitué par Dominique Labbé à Grenoble). (Habert et al, 1997 : p. 18)

soi qu'il s'agisse d'en choisir une au sein d'une même famille, d'identifier les outils disponibles, de lire et interpréter les données, sans parler de l'utilisation des logiciels.

Outre la lourdeur des traitements, on constate également une difficulté à interpréter les résultats qui dépendent grandement des méthodes utilisées pour « mesurer » des proximités. Etienne Brunet montre que lorsqu'il s'agit d'apprécier la distance intertextuelle, les problèmes sont de plusieurs ordres (Brunet, 2003) : d'une part, sur quels éléments comptables faut-il s'appuyer ? Ce qui revient à se demander ce que l'on mesure exactement, et d'autre part, quelles méthodes il faut utiliser et comment interpréter les résultats. De nombreux travaux de recherche mobilisent des logiciels d'analyse de contenu ou de discours sans se préoccuper nullement de ces éléments de questionnements. Or, Etienne Brunet indique à juste titre :

> *« La distance entre deux textes, c'est comme la distance entre deux êtres ou entre deux cultures. Il ne semble pas qu'on puisse appliquer là la mesure. […]Certes le texte a une réalité matérielle qui se prête à l'analyse. Les éléments comptables peuvent être soumis aux instruments de mesure, des plus menus (les atomes des lettres) aux plus volumineux (les grosses molécules des structures syntaxiques, des schémas narratifs ou topoi, des constellations lexicales ou thématiques). Mais il y a tant de tests, tant de mesures, d'indices, de dosages et de stylogrammes que le jugement reste en suspens. On sort perplexe du laboratoire, comme on sort perplexe, cardiogrammes sous le bras, après un bilan de santé. » (Brunet, 2003 : p. 2)*

En ce sens, nous avons pu constater que les méthodes développées par la fouille de textes (classification supervisée, non supervisée) ne prennent pas en compte les variations textuelles fines contrairement aux méthodes issues des statistiques textuelles qui se sont développées depuis les travaux précurseurs de Pierre Guiraud en 1954, Charles Muller en 1973, puis les méthodes d'analyse de données de Jean-Paul Benzécri en 1973. Bien que les deux familles de méthodes s'attachent à traiter de manière robuste de grands ensembles de données, les méthodes de classification sont plus proches des méthodes informationnelles (identifier des termes pertinents sur des principes statistiques et en recourant par exemple à des mesures d'information mutuelle qui permettent de prendre en compte les associations de termes), alors que les méthodes issues des statistiques textuelles sont plus enclines à

donner un sens linguistique aux calculs mathématiques[105] : prise en compte des champs thématiques (cf. les études conduites avec le logiciel Hyperbase sur les textes littéraires), ou la prise en compte des spécificités, l'évolution de la richesse lexicale, etc. De ce fait, la lecture des résultats de la classification automatique n'est pas toujours aisée et l'expertise linguistique plus difficile à appliquer sur des nuages de points… Ces méthodes gagneraient à être connues en développant des outils plus accessibles aux « artisans du texte ».

2.3.3. Pour conclure sur l'organisation des connaissances

A ce point de notre travail, la question de l'organisation des connaissances doit être mise en perspective. Dans le premier chapitre, l'enjeu se focalisait sur des connaissances représentées suivant une tradition logico-grammaticale, conduisant à privilégier les opérations de segmentation (des phrases, des propositions, des mots, des morphèmes, etc.) de hiérarchisation (par le recours à des règles contextuelles et hors contexte) et de relations linguistiques les unités identifiées (relations morpho-phonologiques, de syntaxe interne, de sémantique lexicale). Dans le second chapitre, l'organisation des connaissances portait sur une entité complexe, le texte, dont la description était abordée comme une corrélation de traits caractéristiques du genre, entité sociale et linguistique. Cette approche privilégiait une approche empirique des corpus, prenait en considération les normes d'usages appréhendées à travers les fréquences, et privilégiait l'approche différentielle, c'est-à-dire une approche du sens s'intéressant aussi bien aux connaissances en « creux », les connaissances absentes des textes, qu'aux connaissances récurrentes dans les textes.

Dans les deux cas, l'organisation des connaissances est appréhendée en faisant intervenir les niveaux de description du langage pour éclairer les plans d'organisation de l'information. Ces réflexions se situent toutes deux dans le contexte de la recherche d'information ayant pour objectif de faire converger les représentations des requêtes et des documents. La première approche permet de normaliser les contenus en faisant des choix sur les formes canoniques. La seconde approche permet

[105] voir sur ce point la présentation des méthodes de textométrie par Bénédicte Pincemin. [en ligne] http://textometrie.ens-lyon.fr/spip.php?rubrique80

de décrire des textes au moyen de descripteurs caractéristiques du genre textuel, ce qui permettra de donner en réponse à l'utilisateur, des textes de même genre.

De manière symétrique, se pose la question du codage des contenus et de leur formalisation, opérations qui engagent une interprétation ; Bruno Bachimont indique que l'on peut distinguer deux types de formalisations selon que l'on modélise la forme d'expression d'un contenu ou sa signification (Bachimont, 2007). Dans le premier cas, il s'agit d'une *codification* qui va entraîner *« une modification de la forme perceptible »* sans que la signification de cette modification perceptible soit prédictible. Dans le second cas, il s'agit d'une formalisation lorsque la signification est directement appréhendée par la modélisation en négligeant la forme d'expression. Pour résumer, il convient donc de distinguer *« plusieurs niveaux dans le passage du contenu à son équivalent codifié puis formalisé »* (ibid., p. 40) :
Il reste que cette codification n'est pas suffisante pour constituer des connaissances contextualisées :

> *« En effet en considérant qu'il suffit de formaliser les inscriptions pour dégager ce qui constitue en elles la connaissance, le formalisme isole les inscriptions de leur contexte et de leur appartenance à des pratiques culturelles et sociales qui conditionnent leur lecture et utilisation. En faisant ainsi, le formalisme se coupe la possibilité de considérer le sens des inscriptions tel qu'il est et privilégie une reconstruction obtenue par l'axiomatisation de la sémantique des inscriptions » (Bachimont, 2007, p. 60-61)*

La question de la contextualisation, du rapport aux usages et aux pratiques fait l'objet de la seconde partie de ce document.

Deuxième partie. Organisation des connaissances et dispositifs informationnels : vers une mise en contexte

Dans cette seconde partie, nous posons la question de l'organisation des connaissances dans une perspective qui appréhende l'information, sa recherche, sa création, son exploitation, etc. en lien avec les pratiques sociales. Ce changement de perspective nous conduit à repenser les méthodes d'analyse et de traitement des informations, à les envisager au filtre des pratiques informationnelles, des documents et des dispositifs. Nous nous plaçons donc dans une perspective contextualisée de la recherche d'information. En effet, contrairement à la première partie, nous n'envisageons plus la recherche d'information comme un ensemble de techniques et de méthodes permettant « le retrouvage d'information » ainsi qu'il est d'usage dans le courant de *l'information retrieval* mais comme une pluralité d'activités informationnelles et communicationnelles, de pratiques et d'usages de l'information, réalisées par les individus grâce à des dispositifs informationnels.

Cette conception de la recherche d'information n'est pas nouvelle. Les premiers travaux centrés sur les modèles utilisateurs remontent aux années 70, mais c'est à partir des années 90 que ces modèles ont trouvé davantage d'écho auprès du courant de l'*information retrieval*, essentiellement en raison des critiques qui ont été émises sur les protocoles d'évaluation des performances des systèmes définis en laboratoire (*laboratory-based evaluation*). Deux grandes familles de modèles accordent à l'usager[106] une place centrale dans la modélisation de l'activité de recherche d'information : les modèles « orientés utilisateurs » et les modèles « orientés usagers » qui présentent la particularité d'intégrer des facteurs cognitifs et sociaux liés à la recherche d'information vue comme un « processus » ou comme des « activités humaines ». Deux ouvrages donnent à voir l'état de l'art de la littérature anglo-saxonne sur les modèles qui prennent en considération les facteurs cognitifs, sociaux et contextuels (Ingwersen et Järvelin, 2005 ; Case, 2002) ; en français, nous

[106] En introduction de leur ouvrage, Nicole Boubée et André Tricot, notent que « *le terme usager est bien souvent le vocable employé dans les études consacrées à l'étude des interactions avec les systèmes d'informations* ». En outre, les auteurs indiquent que lorsque « *l'activité de l'usager passe au premier plan des théorisations, l'expression chercheur d'information […] est appropriée* » (Boubée et Tricot, 2010 : p. 8).

citerons notamment les ouvrages (Chaudiron et Ihadjadene, 2004a ; Boubée et Tricot, 2010) ainsi que différents articles de synthèse (Ihadjadene et Chaudiron 2008 ; Ihadjadene et Chaudiron, 2010 ; Maurel, 2010), les mémoires d'habilitation à diriger les recherches de Vincent Liquète (2011) et de Céline Paganelli (2012).

Les terminologies francophone et anglo-saxonne sont révélatrices de la manière dont a été prise en compte la dimension humaine dans les activités de recherche d'information. Au sein de la Library and Information Science (LIS), sont en usage les expressions « *information seeking* », « *information searching* », « *information behaviour* » et « *information practice* » qui considèrent l'activité humaine de recherche d'information en contexte. Ces termes ne sont pas interchangeables, ils révèlent des focales particulières et se distinguent par l'étendue du contexte considéré (Ihadjadene et Chaudiron, 2008 : p. 185). « *Information seeking* » décrit le processus de recherche d'information dans son ensemble. Il prend en compte le déroulement chronologique des étapes de recherche d'information initiée par le « besoin d'information » ainsi que divers éléments situationnels qui interviennent avant/pendant/après la recherche d'information. L'expression « *information searching* » décrit plus précisément l'interaction entre l'individu et le système et se focalise sur la formulation des requêtes. « *Information behaviour* » décrit l'ensemble des comportements informationnels des individus ou groupes d'individus en situation de recherche d'information. Cette expression peut décrire l'information intentionnelle (*information seeking behavior*) ou les actions physiques ou mentales associées utiles à l'appropriation de l'information trouvée (*information use behaviour*) (Boubée et Tricot, 2010 p. 19-20). Enfin, « *information practice* » désigne selon Reijo Savolainen « *un ensemble de moyens établis socialement et culturellement permettant d'identifier, de rechercher, d'utiliser et de partager l'information disponible dans différentes sources telles que la télévision, les journaux et l'Internet* »[107]. Cette conception sociale de la recherche d'information ancrée dans le quotidien cherche à se démarquer de l'approche cognitive « *information behaviour* », qui appréhende les pratiques informationnelles comme gouvernées par des « *besoins individuels* »

[107] Information practice *« may be understood as a set of socially and culturally established ways to identify,seek and use, and share the information available in various sources such as television, newspapers, and the internet »* (Salovainen, 2008 : p. 2)

(Savolainen, 2008). En français, l'expression « *pratiques informationnelles* » est la traduction pour « *information behaviour* » et « *information practice* ». Cette traduction ne permet pas de distinguer les deux plans d'analyse psychologique *versus* sociologique. Les termes de « comportements informationnels » pour « *information behaviour* » et d'«activités informationnelles » pour « *information in social practice* » (Cox, 2008 : p. 185) permettent de rendre compte de ces différences. Nous avons adopté l'expression « dispositifs informationnels » ou encore « dispositifs d'accès à l'information » (Simonnot, 2012) en lieu et place de « système de recherche d'information » Ce changement de terminologie traduit un changement de paradigme. Les dispositifs d'accès à l'information désignent « *l'ensemble des lieux et des objets de médiation, techniques ou non, permettant d'identifier, de repérer et de collecter et éventuellement traiter l'information* » (Ihadjadene et Chaudiron, 2008), également (Couzinet, 2009, 2011), et pour Brigitte Simonnot, « *ils n'ont pas de sens sans leurs usagers* » (Simonnot, 2012). Dans ce cadre ouvert, la recherche d'information n'est pas uniquement médiatisée par des techniques, la communication verbale pouvant aussi permettre de « s'informer ». D'après Brigitte Guyot, chercheur en SIC, ces dispositifs informationnels sont des objets socialement construits appréhendés dans « *leur interrelation avec les acteurs (concepteurs et utilisateurs), les objets (l'information et les outils) à travers une interface (qui peut éventuellement être un lieu), favorisant ainsi la création de représentations diverses* » (Guyot, 2006).

Dès lors, nous élargissons notre cadre d'observation. Après avoir analysé les systèmes dans leur complexité interne, nous envisageons le contexte social dans lequel les dispositifs prennent place. Ceci nous amène à considérer les individus qui font usage des dispositifs, leurs pratiques informationnelles et communicationnelles, l'environnement social dans lequel ils se situent, les tâches qu'ils ont à accomplir. Cette complexité n'est pas catégorisée en tant que telle dans un modèle holistique de la recherche d'information. Elle contribue à identifier les éléments qui jouent un rôle sur l'organisation des connaissances. Si le vocabulaire a longtemps été la question centrale de l'organisation des connaissances, bien d'autres dimensions sont à prendre en compte : les modes de sélection et de consultation des sources d'information, les choix de parcours de lecture, les formes d'annotation des documents, l'exploitation qui est faite de l'information, etc. L'observation de ces pratiques informationnelles et

communicationnelles permet de distinguer différents niveaux d'information propres à faire émerger des connaissances. Comment alors identifier ces connaissances ? Comment sont-elles organisées ? Quelles relations existe-t-il entre elles ? Ainsi que nous l'annoncions dans l'introduction de ce mémoire, nous considérons que les connaissances sont le fruit d'une construction théorique et méthodologique, *in fine*, elles sont formalisées, ce qui est du ressort de l'ingénierie des connaissances. Sur le plan théorique, les connaissances sont issues d'une confrontation entre divers niveaux d'information mêlant la technique, le langage et le social[108]. Sur le plan méthodologique, les discours constituent le matériau d'analyse de l'information. Ils sont recueillis dans le cadre d'une recherche d'information au sens défini *supra* et font intervenir les représentations formalisées dans les dispositifs, les contenus édités, les commentaires d'usagers sur leurs popres pratiques informationnelles[109] **Cette approche peut être qualifiée d'immersive, puisqu'elle permet de saisir dans les discours produits lors d'une recherche d'information, des informations propres à identifier des connaissances.**

Cette partie comporte trois chapitres. Le premier chapitre aborde l'environnement immédiat de notre recherche et dresse un état des lieux de la manière dont le contexte est abordé en recherche d'information. Le second chapitre présente les méthodes de constitution de corpus pour analyser les discours et le troisième chapitre, envisage l'apport de notre travail à l'organisation des connaissances. Le nombre de publications qui portent sur ce thème est de 16 sur 45 qui ont été publiées entre 2007 et 2013.

108 *« Pour nous, le monde social fait référence aux composantes de la réalité qui ont une existence objective mais qui sont le résultat de l'activité humaine. À ce titre, des réalités aussi diverses que : l'écriture, l'argent, les États, les villes, le football… et, bien entendu, les techniques font partie du monde social. […] Précisons également que l'expression « ancrage dans le monde social » ne devrait pas être interprétée comme étant une « insertion » des techniques dans le monde social, car elles en font partie au même titre que les autres composantes de cette réalité. Il s'agit donc d'interdépendances entre réalités ayant un statut ontologique similaire ; à ce niveau, l'expression « ancrage dans » désigne donc tout au plus l'orientation d'un regard qui cherche à penser cette articulation à partir de l'étude des techniques. »* (Staii, 2012 : p. 88).

109 Bernard Miège considère les dispositifs comme « *une articulation d'outils, de contenus et d'usages* » (Miège, 2007 : p. 47).

1. L'environnement de notre recherche : éléments de cadrage

1.1. Le GRESEC : un cadre de recherche en SIC

En 2004, nous avons obtenu un poste de maître de conférences par mutation à l'UFR des sciences de la communication[110] de l'Université Stendhal, Grenoble 3. Nous sommes devenue membre permanent du GRESEC dirigé par Isabelle Pailliart. Pour la première fois, nous avons effectué notre service d'enseignement dans la même discipline que notre recherche à l'institut de la communication et des médias. En recherche, nous avons été intégrée à l'axe Connaissance, Recherche d'information, Interfaces et Systèmes de Traitement Automatique de la Langue (CRISTAL). Cet axe a été rebaptisé en 2011 Connaissance Information Document (C.I.D) en raison de l'abandon des recherches en informatique, en psychologie cognitive et traitement automatique des langues. Laurence Balicco avait pris la responsabilité de l'axe après le départ à la retraite de Jaques Rouault en 1999. Les thématiques de recherche de C.I.D ont été profondément remaniées au fil des années et ces bouleversements ne peuvent être compris si l'on ne prend en compte quelques éléments de contexte. Nous montrerons ensuite quel a été notre rôle dans cette réorientation.

1.1.1. Les sciences de l'information se dégagent de l'ingénierie

A partir des années 2000, au sein des SIC, se sont redessinés les contours des sciences de l'information[111]. Les membres de la discipline issus du courant de *l'information retrieval* se sont progressivement dégagés de « l'ingénierie », perspective qui semblait entacher les sciences de l'information de ses lettres de noblesse en recherche[112] (Couzinet, 2002). Mais l'exercice était périlleux, les racines de « la » science de l'information puisant dans les techniques documentaires et constituant l'ingrédient majeur de l'information-documentation. Hubert Fondin a été l'un des chercheurs influents en sciences de l'information à revendiquer une posture épistémologique en

[110] En 2012, l'UFR des sciences de la communication est devenue le département des sciences de l'information et de la communication au sein de l'UFR langage, lettres et arts du spectacle, information et communication (Llasic).

111 Cf. la délimitation du champ des SIC sur : http://www.cpcnu.fr/web/section-71

[112] Viviane Couzinet constate « *qu'en France*, [les sciences de l'information] *apparaissent plus tournées vers l'ingénierie que vers la recherche* ». (Couzinet, 2002 : p. 11).

sciences humaines et sociales (Fondin, 2001 ; 2002). Il mentionne que cette réflexion a été initiée en 1997 avec Jacques Rouault dans un texte diffusé de manière restreinte aux seuls membres du CNU[113].

> *« Le choix épistémologique retenu est celui d'une science humaine et sociale appartenant aux sciences de l'information et de la communication, les SIC, ce qui écarte toute revendication sur un objet physique ou réel. La SI a pour objet scientifique – pour objet « construit » – le processus de recherche d'information. Elle étudie, pour essayer de les comprendre, les modalités – humaines et techniques – de cette recherche vécue comme une rencontre « virtuelle » entre un producteur d'information et un rechercheur d'information, recherche souvent médiatisée par des outils, des spécialistes et des lieux. C'est une étude sur le sens – sens fourni, sens attendu, sens partagé entre des personnes dans le cadre d'un système ou d'un dispositif informationnel. » (Fondin, 2002)*

Au GRESEC, ce mouvement se traduisait par un rapprochement de l'information et de la communication ainsi que le révèle un ouvrage collectif dirigé par Bernard Miège auquel avait participé une quinzaine de chercheurs de différentes disciplines, dont Jacques Rouault. Le titre de cet ouvrage « *Communication personnes systèmes informationnels* » est révélateur d'une tentative de rapprochement de l'information avec les SHS, alors que le dialogue homme-machine était par tradition abordé par l'informatique, la psychologie cognitive, la linguistique et l'informatique. Bernard Miège explique cette nouvelle orientation dans l'introduction :

> *« Nous avons donc entrepris de déconstruire l'objet qui se présentait à nous. C'est ce qui nous a amené par exemple à refuser de situer nos réflexions dans le cadre de la communication homme/machine et à lui substituer l'expression que nous considérons comme plus pertinente (même si elle est sans doute moins heureuse sur le plan éditorial) de communication entre personnes et systèmes informationnels. […] nous entendons élargir ce qui est généralement confiné à la mise au point des logiciels, avec l'appui de la psychologie (ou de l'ergonomie) et de la linguistique ; nous tenons pour préférable une approche où interviennent également la philosophie, les sciences sociales et même la culture. » (Miège, 2003 : p. 18).*

Le chapitre 1 de cet ouvrage est rédigé conjointement par Bernard Miège et Jacques Rouault : les concepts fondamentaux (information, système d'information, interaction avec des personnes, etc.) sont présentés avec un double regard informationnel et

[113] Ce texte était intitulé « L'information : l'Arlésienne de l'interdiscipline des sciences de l'information et de la communication » (Fondin, 2001, note 1)

communicationnel. Les enjeux scientifiques conflictuels qui traversent le domaine de la communication homme-machine sont évoqués, et la terminologie est discutée. Par exemple, les notions d'utilisation, issue des sciences cognitives et d'usage, issue des sciences sociales sont présentées comme complémentaires :

> *« Jusqu'ici nous avons surtout évoqué l'utilisateur pour désigner le protagoniste humain d'une interaction personne-système d'information. Ceci sous-entend que nous cantonnons son étude à sa position par rapport au système d'information ; ce qui va être important, ce sera son habileté à utiliser le système d'information, ainsi que ses connaissances des informations, ou des connaissances stockées. Un autre point de vue sur ces personnes est celui de l'usager : l'interaction est alors étudiée dans ses rapports avec d'autres usagers et avec la société environnante. [… Les contributions de ce volume sont tantôt orientées vers l'utilisateur, tantôt vers les usagers. » (Rouault et Miège, 2003 : p. 28).*

Au sein de l'axe CRISTAL, cette nouvelle posture provoquait une vague de départs des chercheurs en informatique, linguistique et psychologie en sus du départ à la retraite de Jacques Rouault. Ainsi que l'a très justement souligné Adrian Staii, membre de notre axe, il s'est ensuivi un certain « malaise identitaire » de l'information-documentation (Staii, 2004), qui explique peut-être la « discrétion des sciences de l'information » et leur manque de visibilité au sein des SIC (Couzinet, 2002). Ce malaise était bien réel dans l'axe CRISTAL, dont les thèmes de recherche concernaient le traitement et l'utilisation de l'information, la représentation des connaissances dans le champ de l'information spécialisée et professionnelle. Pour traiter ces sujets, différents outils conceptuels et méthodologiques étaient mobilisés puisant dans des disciplines connexes, tels que l'informatique, le traitement automatique des langues, la linguistique de corpus, la psychologie cognitive et le dialogue homme-machine. La question de la représentation des connaissances par exemple, abordée sans « toucher » aux codes – langage naturel, langages artificiels, langages de programmation – conduisait à une perte d'identité forte de l'axe. La plupart des compétences qui faisaient la force de l'axe étaient désormais hors champ et se trouvaient reléguées au niveau de compétences méthodologiques de second degré. Par ailleurs, progressivement, les revues dans lesquelles publiaient traditionnellement les chercheurs de l'axe (*Tal, Document Numérique, Revue d'intelligence artificielle, Revue Interaction homme/machine, Modèles linguistiques, Travail Humain, etc.)* étaient éliminées de la liste des revues préconisées par le CNU…

alors même qu'en 2010, il était encore admis que « le traitement automatique de la langue ou encore les interfaces hommes-machines » figurent comme objets de recherche de la discipline. En 2013, ces thématiques ont fini par disparaître de la définition du champ des SIC et les préconisations sont (presque) cohérentes avec les titres de revues[114]. On peut toutefois déplorer que le nombre de revues dédiées aux sciences de l'information soit moindre que les revues consacrées aux sciences de la communication, cette situation étant liée notamment au faible nombre de chercheurs en sciences de l'information, peut-être aussi à la tradition collective de la recherche.

Progressivement, se sont alors profilées de nouvelles perspectives de recherche prenant en compte le contexte social et technique de l'information-communication : citons, parmi les travaux des membres de l'axe, le rôle des experts dans la recherche d'information (Mounier, 2013), la prise en compte des activités info-communicationnelles en contexte de travail (Paganelli, 2012), ou l'analyse de la nature et de la fonction sociale et communicationnelle des TIC, dont les systèmes de recherche d'information ne sont qu'un artefact parmi d'autres (Staii, 2012).

1.1.2. L'enseignement des sciences de l'information

Un second élément de contextualisation, plus local celui-ci, mérite d'être souligné : il concerne la formation par la recherche qui s'est trouvée progressivement concurrencée par les filières professionnelles. A partir de 2010, les parcours (ou options) consacrés aux sciences de l'information n'ont plus trouvé suffisamment de candidats, et le parcours guidé *« Gestion des connaissances et information spécialisée »* de la licence 3 information-communication, ainsi que l'option « *Information, documents, usages et systèmes* » du master 2 de recherche en sciences de l'information et de la communication ont été supprimés. Nous attribuons la désaffection des candidats à une question d'offre. D'une part, se sont développées des filières professionnelles dans l'université voisine, l'Université Pierre Mendès-France : licences professionnelles à l'IUT2, master professionnel sur les métiers de la documentation et des bibliothèques. A l'Université Stendhal, s'est également

[114] Les revues internationales les plus prestigieuses comme Jasist ou Arist ont une ligne éditoriale très nettement orientée vers *l'information retrieval*, mais il s'agit ensuite d'assumer « l'exception française ».

développé un master très attractif, le master industrie de la langue en sciences du langage. Ces filières ont attiré le (petit) vivier d'étudiants intéressés par les sciences de l'information. D'autre part, le volume horaire consacré à l'information dans le master SIC était devenu trop faible en licence et master par rapport aux enseignements de communication pour attirer des étudiants souhaitant se spécialiser en information-documentation.

Entre les années 2004 et 2011, l'introduction des sciences de l'information dans le cœur de formation de la licence et du master 1 et 2 s'est traduite par un effort important d'adaptation aux exigences de ces filières, conduisant par exemple à éliminer certains cours « techniques » que nous-même et les collègues de l'axe[115] dispensions : par exemple, le TAL, la programmation en Prolog, les cours de gestion documentaire, les cours sur les interfaces, etc. Nous avons redéfini des cours plus généraux en sciences de l'information (dès le niveau L1), maintenu des cours qui permettent de « réfléchir la technique » ou d'appréhender la question de la représentation des connaissances dans le cadre d'études nécessitant de manipuler des outils d'annotation issus du TAL et de standard XML. Nous avons mis l'accent sur des enseignements en lien avec les usages et les pratiques informationnelles. Enfin, nous avons maintenu des cours plus spécialisés en les articulant avec des enseignements méthodologiques complémentaires de ceux dispensés en communication. Par exemple, en 2011, Fabienne Martin-Juchat et nous-même avons mis en place un cours de bibliographie anglo-saxonne destiné aux étudiants de master 1 Recherche et Etudes en Information et Communication (RETIC). Cet enseignement s'appuie sur des fondamentaux en sciences de la communication – le lien est assuré par notre collègue – et en sciences de l'information, en lien avec notre cours Information spécialisée.

Bien que nous ne soyons pas le seul artisan de cette intégration[116], la dénomination institutionnelle est révélatrice de cette évolution : *l'UFR des sciences de la*

[115] Ils étaient tous (et pour certains le sont encore) enseignants à l'Université Pierre Mendès-France.

[116] En 2010-2011, nous étions membre du bureau de l'UFR des sciences de la communication — dont la direction était assurée par Fabienne Martin-Juchat — au moment où s'est faite cette restructuration et nous avons participé à la modification des textes du futur département.

communication est devenue en 2012, lors de la restructuration de l'Université, le *département des sciences de l'information et de la communication*, dénomination qui s'est imposée logiquement. Le point sombre réside dans le constat attristé du faible nombre d'étudiants et de chercheurs attirés par les sciences de l'information, ce qui nous semble dû en partie à un déficit de formation des étudiants dans les premiers cycles, ce retard étant difficilement rattrapable ultérieurement.

1.1.3. Notre participation à des recherches finalisées

A notre arrivée dans l'axe CRISTAL, nous avons fait le constat que les recherches en TAL avaient « raté le coche » du TAL robuste. Nous avons proposé divers séminaires de formation sur cette thématique[117] mais sans grand succès : la demande sociale sur les études d'usage et de pratiques se faisait de plus en plus pressante. C'est ainsi que laissant temporairement de côté nos préoccupations liées à la caractérisation des contenus textuels, nous avons participé à plusieurs études empiriques, la plupart situées dans le cadre de recherches finalisées.

1.1.3.1. Le projet NOESIS : Une plate-forme d'aide à la recherche d'information médicale

En 2006, le centre hospitalier universitaire de Grenoble (CHUG) avait sollicité le GRESEC pour conduire une étude sur les pratiques informationnelles des médecins de l'hôpital. Cette étude faisait partie d'un programme de recherche européen (NOESIS) destiné à développer un module d'aide à la recherche d'information médicale sur une plate-forme de « *knowledge management* ». Cette étude nous a permis de nous former aux méthodes des entretiens, et nous a ouvert un champ de recherche nouveau mettant en avant le contexte de recherche professionnel

[117] Intervention au séminaire de recherche doctorants, interne à l'axe Cristal en février 2005 : « Outils et méthodes de traitement de corpus » avec Céline Poudat, membre du Coral de l'Université d'Orléans et de Modyco de l'université Paris 10, Février 2005.
Intervention au séminaire de recherche doctorants, interne à l'axe Cristal en mars 2007 : *Le genre comme point d'accès au document : analyse comparée de textes en linguistique et mécanique,* en février 2007.
Intervention au séminaire de recherche doctorants, interne à l'axe Cristal du Gresec « *Analyse de discours pour le TAL et recherche d'information* » avec Hélène Braoudakis, doctorante au Gresec, janvier 2008.

« quotidien » des médecins. La recherche s'est déroulée de manière collective[118] et a donné lieu à un rapport de recherche (Balicco et al., 2007) et un article collectif (Staii et al., 2008). Grâce à cette recherche, nous avons découvert que dans le courant de la *Library and Information Sciences* (LIS), la santé était un domaine de la recherche d'information à part entière (*Health Information seeking and behaviour*) dont les enjeux se conçoivent au niveau des patients ou des médecins et qui se cristallisent autour de la prise de décision. Après avoir dressé un état de l'art sur la médecine fondée sur les preuves, ou médecine factuelle (*Evidence-based medicine*), participé à l'analyse des entretiens et contribué à l'identification des sources d'informations utiles aux médecins interrogés, nous avons pris conscience du rôle de l'information dans les pratiques professionnelles.

1.1.3.2. L'ANR SCIENTEXT : « un corpus et des outils pour étudier le positionnement et le raisonnement de l'auteur dans les écrits scientifiques. »

En 2007, nous avons participé à l'ANR « *Scientext. Un corpus et des outils pour étudier le positionnement et le raisonnement de l'auteur dans les écrits scientifiques* », dirigée par Francis Grossmann et Agnès Tutin du laboratoire de sciences du langage Lidilem à l'Université Stendhal. Ce projet dont le descriptif figure sur le site http://scientext.msh-alpes.fr/ *« met à la disposition des chercheurs et étudiants un large corpus d'écrits scientifiques de manière à permettre l'étude de leurs caractéristiques linguistiques.* ». Ce corpus a été constitué dans la perspective d'étudier le positionnement et le raisonnement à travers la phraséologie, les marques énonciatives et les marques syntaxiques liées à la causalité.

Nous avons participé à des séminaires et des journées d'études et de recherche en 2008 et en 2010 et avons continué à nous former aux techniques et méthodes d'analyse de corpus[119]. Nous avons mis à jour nos connaissances sur les dernières évolutions de la TEI qui venait de passer à sa version P5. En contrepartie, nous avons

[118] Tous les membres de l'axe y ont participé : Laurence Balicco, Marc Bertier, Céline Paganelli, Evelyne Mounier, Adrian Staii et moi-même.

[119] Notamment le logiciel Nooj que nous a fait découvrir Agnès Tutin.

contribué à alimenter le corpus scientifique dans le domaine de la mécanique grâce à l'aide d'Annie Leroy-Chesneau, enseignant-chercheur au laboratoire Prisme de l'Université d'Orléans, qui a su convaincre ses collègues de nous transmettre leurs mémoires de thèse et d'HDR. Puis, en collaboration avec Céline Paganelli, nous avons abordé la notion de positionnement scientifique du point de vue des sciences de l'information. Cette collaboration a été très importante pour notre activité scientifique, elle a donné lieu à plusieurs publications (Clavier et Paganelli, 2010, 2012a et b). Ces travaux participent de l'observation des pratiques informationnelles et de la caractérisation de contenus en milieu professionnel.

1.1.3.3. Les projets CaNu XIX et Métilde : le patrimoine numérique du 19ème siècle

En 2008-2010 puis en 2011-2014, nous nous sommes impliquée dans deux recherches commanditées ayant trait à la valorisation du patrimoine numérique du 19ème.

Le premier projet financé par la région Rhône-Alpes s'intitulait CaNu XIX, pour Canards Numériques du 19e siècle. Déposé en 2007 par Geneviève Lallich-Boidin du laboratoire lyonnais ELICCO, CaNu XIX portait sur la valorisation et la mise en ligne de fonds patrimoniaux de la presse illustrée du 19ème siècle. Nous avons participé à la thématique 5 de ce programme de recherche intitulé « Vers une construction des langages d'indexation et de recherche ». Il s'agissait de « réfléchir à une automatisation partielle de l'indexation […] en suggérant des zones de documents susceptibles de contenir des informations pertinentes, des indices linguistiques, des unités d'information porteuses de sens, des indications sur les connaissances contenues dans les documents ». Dans CaNu XIX, nous avons montré que les parcours thématiques sont une forme d'accès au patrimoine de presse à développer pour le grand public et que les événements sont des informations de toute première importance dans ce type de sources (Clavier, 2010).

Le second projet, intitulé Métilde du nom de l'un des premiers grands amours de Stendhal, est une structure fédérative de recherche interne financée par l'Université

portant sur la mise en ligne des manuscrits de Stendhal sur une plate-forme documentaire dénommée Clélia. Métilde est sous la responsabilité de Cécile Meynard du laboratoire Traverses 19-21 en collaboration avec Thomas Lebarbé du Lidilem. Nous assurons la responsabilité de la partie du projet relevant des sciences de l'information et la communication et qui comporte trois volets : le premier porte sur l'élargissement des publics (spécialistes et grand public) utilisant la plate-forme documentaire ; le second volet consiste à analyser les formes de porosité entre la médiation technique et humaine ; le dernier volet est lié à la valorisation de la collection par des propositions de parcours. Ce projet est en cours de réalisation, nous travaillons à la coordination des différentes tâches conduites par les 7 chercheurs en SIC et participons à certaines d'entre elles.

1.1.3.4. Les études sur le thème de la santé : cancer et autres maladies

Hormis ces recherches finalisées, nous avons participé à diverses études dans le domaine de la santé. Ces études ont permis d'établir des collaborations en interne avec d'autres chercheurs du GRESEC et d'ouvrir les thématiques de l'axe sur l'information grand public, alors que jusqu'à présent elles se limitaient à l'information spécialisée. Trois études ont été réalisées, l'une en collaboration avec Hélène Romeyer sur l'analyse des discours sur le cancer (Clavier et Romeyer, 2008) ; la deuxième est une recherche collective sur l'analyse des discours dans le forum de discussion Doctissimo (Clavier et al. 2010) et la troisième, constitue une étude plus pointue que la précédente portant sur les maladies rares et orphelines dans Doctissimo, et qui aborde les forums comme des ressources informationnelles hybrides. Cette dernière étude a été conduite en collaboration avec Céline Paganelli (Paganelli et Clavier, 2011).

Dans la continuité de ces travaux, nous avons fait une demande d'allocation à diriger les recherches auprès de la région Rhône-Alpes en 2009. Cette allocation a été acceptée en 2010, nous avons pu participer à l'encadrement de Céline Battaïa en sciences de l'information et de la communication dont la thèse a été soutenue le 10 décembre 2013. Le titre du mémoire est « *L'émotion dans les forums de discussion : structuration et évaluation de l'information de santé* » (Battaïa, 2013).

1.2.4. Notre contribution à un programme de recherche dans l'axe C.I.D

En 2011, l'axe CRISTAL a été renommé Connaissance, Information et Document (C.I.D) et l'orientation sur l'étude des usages et des pratiques informationnelles s'est affirmée en lien avec les thématiques développées dans l'axe « Ancrage social des techniques en information-communication » dont Adrian Staii venait de prendre la responsabilité. Pour le contrat quinquennal 2011-2016, Céline Paganelli et moi-même avons proposé un programme de recherche dont le titre n'était pas très heureux (*Recherche d'information : modélisation et usages*) mais dont le contenu nous a permis de préciser nos contributions respectives.

Nous avons situé notre recherche dans le champ de « *l'information spécialisée* » et avons fait état de deux ensembles de travaux réalisés dans le passé : les premiers avaient montré que *« les pratiques des usagers en matière d'accès à l'information étaient influencées notamment par le contexte professionnel et l'expertise des individus* »[120] et les seconds que « *la prise en compte de certains marqueurs linguistiques permettait d'améliorer les méthodes de classification et d'indexation des documents*. » L'objectif de ce programme consistait alors « *à relier ces approches et à identifier les régularités qui émergent tant du côté des usages observés que des propriétés linguistiques de l'information*. » Sur le plan théorique, nous posions plusieurs questions, comme celle de « *la finalité du point de vue scientifique d'accumuler des études d'usages et des analyses de corpus au-delà d'une approche descriptive revendiquée* » Et, sur le plan méthodologique, nous nous interrogions également sur « *l'articulation entre des méthodes issues de différents champs disciplinaires, telles que la sociologie (entretiens) la psychologie cognitive (protocoles expérimentaux) et la linguistique de corpus (méthodes symboliques et numériques)* ». Enfin, nous questionnions « *le passage à l'échelle des méthodes de linguistique de corpus* » notamment dans le cadre de la valorisation de bibliothèques numériques.

[120] Ces citations sont extraites d'un document interne « Bilan du programme 11 » rédigé par Céline Paganelli et moi-même qui a été communiqué à la responsable de l'axe C.I.D et à la directrice du GRESEC pour une évaluation à mi-parcours du contrat quinquennal en mars 2013.

L'habilitation à diriger des recherches de Céline Paganelli constitue un premier apport à ce programme de recherche. Son mémoire porte sur l'accès à l'information, notion qui englobe les activités permettant de parvenir aux informations utiles et de travailler avec elles dans un contexte de travail (Paganelli, 2012 : p. 5). Les notions d'usages, de pratiques et d'activités informationnelles sont au cœur du mémoire d'habilitation. En ce qui concerne le second volet du programme lié à la caractérisation de contenus en fonction des usages et des pratiques informationnelles observés, cet objectif nécessitait de repositionner nos objets d'études (les langages, le texte, l'indexation) dans une perspective sociale et culturelle de la recherche d'information.

En premier lieu, se posait la question de la place du langage naturel dans une telle perspective. Si le langage était un élément central du courant de *l'information retrieval*, il semblait complètement évacué des perspectives sociales. Par exemple, dans l'ouvrage de Schatzki et al. (2001) consacré au « tournant théorique sur les pratiques sociales » *(The practice turn in contemporary theory)*, les auteurs évoquent de nombreuses disciplines concernées par l'analyse des pratiques : la philosophie, la sociologie, l'anthropologie, les études scientifiques et technologiques[121]. Les sciences du langage sont absentes. Plus fondamentalement, certains auteurs reconnaissent depuis longtemps que la linguistique et les sciences de l'information entretiennent des liens étroits, puisque « le langage permet de communiquer l'information » mais que le lien entre ces deux disciplines n'a pas été exploité en raison « du peu de chose que la linguistique a à offrir en matière de sémantique et d'explication de ce qu'est la signification en langage naturel » (Montgomery, 1972) :

> *« In theory, the relationship between linguistics and information sciences is clear indisputable: information science is concerned with all aspects of the communication of information, language is the primary medium of the communciation of information, and linguistics is the study of language as a system for communicating information. In practice however, the relationship between the disciplines of linguistics and information science has not been exploited. […] In the first place, linguistics has had very little to offer in the area of semantics, or*

[121] Cité dans (Cox, 2012 : p. 176)

> *explication of meaning in natural language, and it is this aspect of language which is of most concern to information scientists. » (Montgomery, 1972 : p. 195)*[122] *C'est nous qui soulignons*

Murielle Amar, docteur en SIC et actuellement conservateur des bibliothèques à la Biliothèque nationale de France, considère pourtant que les apports de la linguistique pour une approche discursive de l'indexation sont indéniables, invite à la prudence concernant le statut des objets manipulés par les sciences de l'information dans le cadre de la pratique de l'indexation : *« Ce n'est pas parce que l'indexation manipule, entre autres, des objets de nature linguistique (textes, documents, mots des langages documentaires par exemple) qu'elle manipule des objets linguistiques* » (Amar, 2000 : p. 38). Si des liens se sont noués entre la linguistique et les sciences de l'information, c'est bien parce que ces disciplines – disons plutôt, une certaine linguistique ancrée dans la tradition logico-grammaticale et à visée formalisante –, partageaient des points de vue communs sur une conception « représentationniste » du langage. Celui-ci est vu comme une nomenclature, la signification est supposée représentable sous forme d'ontologies, de vocabulaires, de classifications et les représentations obtenues sont censées épuiser le sens des textes. Ce postulat alimente les démarches d'indexation automatique et de représentation des connaissances et perdure dans la conception du web sémantique de Tim Berners Lee. Mais quelle est la part du social dans cette démarche ? Hormis le fait que dans le meilleur des cas, les connaissances sont puisées dans des textes répondant à des normes sociales, comme les genres textuels ou les domaines ?

En deuxième lieu, notre programme posait la question de la place des traitements automatiques, et en particulier de l'indexation, dans une approche sociale de la recherche d'information. En effet, le contexte d'observation d'activités

---

[122] *« En théorie, la relation entre la linguistique et la science de l'information est clairement indiscutable : la sciences de l'information est concernée par tous les aspects de la communication de l'information, le langage est le médium principal de la communication de l'information, et la linguistique est l'étude du langage comme système pour communiquer l'information. En pratique cependant, la relation entre les disciplines de la linguistique et de la science de l'information n'a pas encore été exploitée. […] Premièrement, la linguistique a eu peu à offrir dans le champ de la sémantique, ou de l'explication de la signification dans le langage naturel, et c'est pourtant cet aspect du langage qui concerne le plus les scientifiques en science de l'information »* (Montgomery, 1972 : p. 195 (notre traduction, c'est nous qui soulignons)

informationnelles socialement situées était incompatible avec une conception techno-centrée de l'indexation vue comme processus « instrumental » et « circulaire » au service de la recherche d'information (Amar, 2000 : p. 26-28). D'autres voix en SIC proposaient de penser l'indexation comme « *phénomène communicationnel, social, interprétatif autant qu'informationnel* » (Kovacs et Timimi, 2006), (Courbières 2002). Dans ce cas, si l'indexation était considérée comme un discours interprétatif, nous pouvions alors distinguer deux instances discursives qui participent de la production d'indices. D'une part, les « agents indexeurs » (Menon, 2013 : p. 93), qui, suivant le champ de l'information et de l'organisation concernés produisent une indexation manuelle et/ou automatique. D'autre part, les « usagers » qui, engagés dans des activités informationnelles, laissent des traces de « lecture-écriture » : prise de notes, annotations sur les documents, mais aussi des actions sur les dispositifs, etc. Ces traces ne sont pas des entrées d'index à proprement parler puisqu'elles ne sont ni codifiées, ni organisées, ni normalisées. En revanche, interprétées à la lumière des pratiques informationnelles, les traces deviennent des indices. Et, c'est à ce titre qu'elles peuvent contribuer à l'organisation des connaissances. Cette posture revenait à appréhender les outils de structuration des connaissances, tels les index, comme des instruments de médiatisation de l'information, et non plus seulement comme des outils de normalisation du sens.

En troisième lieu, notre programme posait la question du statut du texte. Le courant de *l'information retrieval* n'établit pas de distinction entre le texte et le document : seul compte le texte dans lequel se loge l'information. Or, en introduisant les individus dans le spectre d'observation, nous faisions du même coup entrer de nouveaux objets, eux-aussi porteurs d'informations. Ainsi, dans le contexte de la recherche professionnelle, le statut des documents consultés est fondamental : par exemple, un médecin lit des articles de revues spécialisées, rarement des livres scientifiques. En outre, les usagers ne lisent pas les textes comme les machines. Le découpage artificiel du texte par niveaux linguistiques, nécessaire pour opérer des traitements automatiques et associer des connaissances homogènes, volait en éclat. Les traces sélectionnées par les usagers dans les documents apparaissaient à tous les niveaux de la textualité (lexique, propositions et parties de propositions, début ou fin de paragraphes, etc.), y compris ceux qui ressortissent du discours, comme les

organisateurs textuels (Ho-Dac et al. 2012), l'anaphore et la co-référence (Corblin, 1995). Ces traces ne concernaient pas uniquement le texte, mais des schémas et des graphiques, etc. Comment alors utiliser ces traces pour documenter les index ? Quelle logique présidait à leur organisation ? Nous avons alors exploré la piste des discours, ces derniers constituent l'élément central du dispositif que nous mettons en place pour recueillir les connaissances. Ces discours sont confrontés aux contenus des documents, et permettent de guider le recueil de connaissances.

Cette articulation entre la caractérisation de contenus et les pratiques informationnelles s'est faite progressivement, et seules les publications portant sur le positionnement de l'auteur sont actuellement les plus abouties Ce changement de posture nous a conduite à revisiter progressivement nos méthodologies en tenant compte du contexte, une notion omniprésente en SIC.

1.2. Le contexte dans les courants de la recherche d'information

Parmi les perspectives qui se profilent dans chacune des deux grandes familles d'approche, technique *versus* humaine, la recherche d'information en contexte se présente comme « la nouvelle perspective », « constitue un enjeu majeur » et « devrait connaître un fort développement » (Chiaramella et Mulhem, 2007 : p. 27). Suivant Peter Ingwersen et Kalervo Järvelin (2005), l'attention portée au contexte est l'élément central d'un « tournant » qu'il faut comprendre comme une tentative d'intégration des deux champs de recherche autour de la notion de contexte donnant lieu à un modèle holistique de la recherche d'information :

> *« Research in Information Seeking and Information Retrieval constitute two disparate research areas or camps. Generally, Information Seeking is rooted in Social Science, with a background in Library Science whereas much of IR is based on Computer Science approaches. The two camps do not communicate much with each other and it is safe to say, that one camp generally views the other as too narrowly bound with technology whereas the other regards the former as an unusable exercise. […] We believe that both research areas can be, and should be, extend to capture more of each other and of context. Therefore, this book seeks to integrate Information Seeking and Information Retrieval into IS&R. » (Ingwersen and Järvelin, 2005 : p. 2)*

Dans la suite, nous envisageons comment les sciences sociales, la linguistique et l'informatique s'emparent de la notion de contexte. Nous laissons de côté les approches cognitives, en particulier la notion de tâches et de cognition située, etc. notre recherche ne nous ayant pas amenée à traiter ces objets scientifiques.

1.2.1. Le contexte pour les approches sociales

Dans le courant de l'« *information seeking* », les modèles contextuels de Tom Wilson qui se sont succédé depuis les années 80 (1981 ; 1994 ; 1997 ; 1999) sont longuement cités dans les quatre plus grandes revues de la LIS[123] (Bawden, 2006). Ils ont également été beaucoup commentés dans la littérature anglo-saxonne et font même école : la perspective Sheffield. Selon Micheline Beaulieu (2003), les travaux du laboratoire de l'Université de Sheffield, the Department of Information Studies, s'inscrivent dans 40 ans de tradition d'études empiriques : d'abord consacrées aux usagers des bibliothèques (1963-1975), ces études se sont ensuite intéressées aux comportements de divers groupes d'usagers (autres que les publics de bibliothèques) ainsi qu'aux besoins informationnels (1976-1988) ; puis au développement d'un cadre théorique pour les études du comportement (1989-1999). Les modèles de Tom Wilson ont pour objectif de relier les variables aux besoins informationnels, aux usages de l'information et aux comportements qui interviennent au sein du processus de recherche d'information. Certaines variables sont liées directement à la personne – physiologiques, psychologiques, émotionnelles –, d'autres sont liées au micro- ou au macro-environnement, d'autres à la position sociale et aux rôles que tient l'usager dans la « vraie vie » (McKenzie, 2003) ou en situation professionnelle[124]. Tom Wilson se réclame d'Alfred Schutz qui a proposé une théorie sociale phénoménologique ayant inspiré les courants ethnométhodologiques et interactionnistes américains (Wilson, 2002).

123 *Journal of the American Society for Information Science and Technology*, *Journal of, Documentation*, *Information Processing and Management*, and *Library and Information Science Research* (Bawden, 2006 : p. 672).

124 *« In practice, context in INS studies usually refers to any factors or variables that are seen to affect individuals' information-seeking behavior: socio-economic conditions, work roles, tasks, problem situations, communities and organizations with their structures and cultures, etc. »* (Talja et al. 1999 : p. 752)

De nombreuses études ont été conduites dans cette perspective, et bien que cette approche souligne que les besoins d'information, la recherche et l'utilisation d'information soient situationnels, la plupart d'entre elles visent néanmoins à trouver des lois universelles ou des modèles de comportement. Cette perspective relève d'une approche objectivée du contexte (*objectified context*), ce qui signifie que les facteurs sociaux, culturels, personnels, situationnels et organisationnels sont conceptualisés comme des entités distinctes et séparées (variables dépendantes et indépendantes) qui contraignent et motivent le comportement des individus de diverses manières (Talja et al, 1999). D'après Maxine Reneker (1993) cité par (Talja et al. 1999) toutes les études conduites dans cette perspective se sont davantage intéressées à l'influence des variables sur les usages et les comportements qu'aux variables elles-mêmes, si bien qu'elles ont failli dans leur mission d'identification des éléments du contexte[125]. Car, même si la désignation des variables permet d'objectiver les entités observées, et d'établir un lien avec la réalité, il n'en reste pas moins que leur caractérisation est loin d'être naturelle. Sanna Talja et al. (1999) montrent que la plupart des variables sont des objets en perpétuelle négociation au sein des disciplines elles-mêmes : comment définir par exemple une structure organisationnelle ?

Dans le courant de « *l'information seeking* », une approche alternative s'est également développée sous le nom de « *sense-making* » qui, selon Dominique Maurel (2010) représente l'un des modèles théoriques les plus utilisés en sciences de l'information pour étudier les comportements informationnels. Initié par Brenda Dervin (1983) qui s'intéresse surtout à la perspective cognitive individuelle, puis appliqué au contexte organisationnel et prenant en considération les groupes (Weick, 1995 ; 2001), le modèle « *sense-making* » relève d'une approche socio-constructiviste qualifiée de modèle « interprétativiste ». Ce type d'approche ne s'intéresse pas à la réalité (i.e au contexte) en tant que facteur extérieur à l'individu, mais au processus subjectif de construction et d'interprétation de la réalité analysé à travers un processus de

125 « *According to Reneker, researchers have failed to correctly identify the variables affecting information seeking and use. It is easy to agree with Reneker's critique, since it is a widely shared notion that the aim of INS studies is to build models of information behavior which show how different factors or variables influence information seeking.* » (Talja et al. 1999 : p. 753)

communication et d'interprétation de l'information (Maurel, 2010 : p. 4). Ainsi, les données recueillies (observations, entretiens, carnets, etc.) présentent-elles différents contextes d'interaction et de construction du sens mais en aucun cas, ne sont considérées comme une réalité figée et continue. Ce sont les acteurs eux-mêmes qui en choisissant de sélectionner tel indice ou signal en provenance de l'environnement vont privilégier une interprétation, la mémoriser, la réactiver suivant leur expérience individuelle, suivant leurs connaissances et leurs émotions, suivant le temps et l'espace. **Par conséquent, comme toutes les activités humaines, la recherche d'information ne consiste pas seulement en des comportements, mais aussi en des significations et des valeurs que les individus accordent à leurs pratiques informationnelles.**

Dans leur ouvrage consacré à la recherche d'information, Nicole Boubée et André Tricot soulignent la diversité des propositions théoriques et empiriques sur la recherche d'information en contexte (Boubée et Tricot, 2010 : p. 166-167). Ils mentionnent qu'il y a deux significations à la notion de contexte : l'une s'appuie sur la dimension sociale du contexte, où le temps et l'espace figurent comme des composantes majeures de la situation ; l'autre s'appuie sur la dimension cognitive, dans laquelle la notion de tâche est primordiale.

1.2.2. Le contexte pour la linguistique

Le contexte compte également pour la linguistique, « la seule discipline contemporaine à placer le réel hors de son objet » et qui « doit sans doute son externalisation du *réel* à l'antique péjoration platonicienne du langage » (Rastier, 1998 : p. 98). Pour faire le lien avec la recherche d'information envisagée comme processus automatisé, nous évoquerons les travaux de Noam Chomsky sur les grammaires formelles qui attribuent au contexte un statut de contrainte (Chomsy, 1957). Ces grammaires conçues pour décrire la compétence des sujets parlants se répartissent en quatre classes de grammaires (nommées de type 0 à 3[126]) qui engendrent des langages pouvant être caractérisés dans le cadre de la théorie des automates. Les langages engendrés par ces grammaires peuvent être soumis ou non

126 Le type 0 est le plus général.

au contexte, ce qui renvoie à la prise en considération (ou non) de contraintes situées à la gauche et à la droite d'un symbole non terminal lors de l'application de règles de réécriture. Pour les grammaires indépendantes du contexte, la réécriture de règles n'est soumise à aucune contrainte contextuelle, mais ce type de grammaire est limité et ne peut pas traiter certains phénomènes langagiers[127]. C'est le cas inverse pour la réécriture des règles dans le cadre des grammaires contextuelles mais leur mise en œuvre est extrêmement lourde, voire impossible[128].

Pour François Rastier (1998), l'abandon des grammaires indépendantes du contexte a été le signe d'un changement de paradigme en sciences du langage. Le contexte permet ainsi d'opposer la tradition logico-grammaticale centrée sur le signe à la tradition rhétorico-herméneutique centrée sur le texte. Suivant la première tradition, le contexte linguistique est considéré comme une « zone d'extension relativement au signe » (*ibid.,* p. 99), le contexte venant éclairer la signification d'une occurrence : il permet la désambiguïsation *versus* la déformation d'un type. Suivant la seconde tradition en revanche, le contexte linguistique est considéré comme une « zone de restriction relativement au texte », le contexte devant alors être pensé en termes de normes – genres, styles – pour interpréter « un passage de texte ». Dans les approches linguistiques contemporaines, le contexte instaure deux rapports différents au réel ce qui est à l'origine des théories de la référence *versus* des théories de l'énonciation. Ainsi que le précise cet auteur, dans les théories de la référence, « les approches formelles conçoivent les signes comme des unités discrètes, des symboles, faisant l'objet d'opérations logiques, et le rôle du contexte est alors de spécifier une occurrence en discours – relativement à son type – défini en langue ». Dans les théories de l'énonciation, le contexte « est décrit comme rapport à la situation d'énonciation qui tient alors lieu de rapport au *réel »* (*ibid.*, p. 98) La situation d'énonciation est définie *hic et nunc*, et peut être considérée comme « une occurrence d'une pratique sociale » (*ibid.*, p. 99). Dans les deux cas, le contexte joue un rôle de

127 Il y a sur-hiérarchisation des structures syntaxiques alors qu'il est parfois nécessaire de traiter des structures plates (coordination). Elles se révèlent par exemple incapables de traiter des constituants discontinus (négation, accord) ou des contraintes à distance entre les constituants, etc.

128 Il est nécessaire d'attribuer des traits sémantiques au vocabulaire terminal pour contraindre les règles de réécriture.

« révélateur », ayant une fonction de sélection – d'une catégorie, d'un trait sémantique, etc. – ou une fonction de localisation – par rapport à un événement extérieur. Par conséquent, il se définit comme un « voisinage local »[129] (p. 106), ou comme une situation. Par tradition, les sciences de l'information peuvent être rattachées à la perspective logico-grammaticale, les sciences de la communication à la perspective rhétorique.[130].

1.2.3. Le contexte pour l'informatique

Pour terminer, la notion de contexte est aussi traitée par les chercheurs en informatique depuis le début des années 2000 (Finkelstein et *al.* 2002 ; Aguiar et Beigbeder, 2004 ; Chiaramella et Mulhem, 2007 ; Hubert, 2010 ; Grivel, 2011 ; Bellot et *al*., 2012 ; Coutaz et *al.,* 2012). Les références bibliographiques relatives au contexte qui sont citées par les informaticiens sont héritées des modèles holistiques tels celui de (Saracevic, 1997) ou (Vakkari et al., 1997) revisités à l'aune de la recherche d'information au sens de (Van Rijsbergen, 1979). Dans ce cadre, le contexte est envisagé comme un ensemble d'éléments à identifier et à modéliser pour améliorer les performances d'un système de recherche d'information. Les éléments contextuels peuvent se rapporter à l'utilisateur – ses connaissances, la tâche, les intentions –, à l'information – le domaine, la structure –, à l'environnement et aux caractéristiques du système (Hubert, 2010 : p. 2). Cette perspective situe explicitement le contexte dans une approche « objectivée », alors que pour les approches sociales, ces dimensions peuvent être appréhendées de manière subjective. Certains modèles se révèlent ambitieux quant à la nature et au nombre d'éléments contextuels qu'ils souhaitent modéliser, et donnent presque l'impression de vouloir modéliser les sciences humaines et sociales dans leur ensemble. Par exemple, l'équipe toulousaine Systèmes d'Informations Généralisés dirigée par Mohand Boughanem[131] propose de distinguer les spécificités internes à l'information, comme le texte, la structure, l'opinion, le domaine ou encore des facteurs externes

129 On parle également de « cotexte ».

130 Ces notions sont abordées dans notre cours *Théorie des écritures*.

131 Présentation de l'équipe SIG/RFI à l'IRIT (Toulouse). http://www.irit.fr/SIG_RFI/

tels que le contexte de la recherche, lequel intègre l'utilisateur, son environnement social, la distance sociale, sa géo-localisation, *etc.*

Devant l'inflation du terme dans la littérature dénoncée d'ailleurs par Lev Finkelstein et al. (2002)[132], ces chercheurs en informatique dressent un panorama des applications et des méthodes qui font intervenir le contexte, notamment pour la recherche d'information sur le web. Le contexte intervient dans divers outils pour aider à l'interprétation d'un contenu suivant une approche guidée par le contexte (*context-driven approach*) pour orienter la navigation ou pour personnaliser les outils. Reposant sur le principe du contexte linguistique évoqué ci-dessus suivant lequel le « voisinage local » permet d'interpréter les contenus, le contexte se rapporte à un domaine ou aux liens hypertextes liés à une page web (Aguiar et Beigbeder, 2004), ou au contexte verbal immédiat (le cotexte), etc. Lynda Tamine et Sylvie Calabretto (2008) justifient les raisons pour lesquelles il est nécessaire de recourir à un traitement contextualisé de l'information :

> *« En clair, le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte d'utilisation particulier. Les besoins sont ainsi énormes. Dans ce cadre, la RI contextuelle émerge comme un domaine à part entière : sans remettre en cause ses origines, elle pose des problématiques nouvelles allant de la modélisation du contexte jusqu'à la modélisation de la pertinence cognitive en passant par la modélisation de l'interaction entre un utilisateur et un système de recherche d'information (SRI). » (Tamine et Calabretto, 2008)*

D'autres travaux portent plus spécifiquement sur le rôle du contexte dans la recherche d'information interactive (Chiaramella et Mulhem, 2007 ; Coutaz et *al.* 2012). Dans ces travaux, l'accent est mis sur « l'utilisateur »[133], le SRI devant s'adapter à l'homme. Cette perspective conduit à développer des interfaces plastiques et adaptables. Ce courant « *observe une démarche holistique selon laquelle le système est envisagé comme un tout pour des usages dans le monde réel avec sa diversité et ses aléas* ». C'est « l'interaction homme-machine systémique ». (Coutaz et al., 2012 : p.

132 « *A large number of recently proposed search enhancement tools have utilized the notion of context, making it one of the most abused terms in the field, referring to a diverse range of ideas from domain-specific search engines to personalization.* » (Finkelstein et al., 2002 : p. 117)

133 En référence notamment au modèle cognitif proposé par Ingwersen (1992) et Ingwersen et Järvelin (2005).

2)[134]. La plasticité des interfaces désigne alors la capacité d'un système interactif à « s'adapter au contexte d'usage […] tout en accordant à l'utilisateur les moyens de contrôle adéquats ». (*ibid*., p. 51)

Nous constatons que c'est encore la recherche d'indicateurs de pertinence adaptés à un contexte d'utilisation de l'information qui justifie ces modèles. Les systèmes d'information peuvent être distingués suivant le nombre de paramètres modélisés, le type de formalisation, la nature des applications envisagées, la manière dont l'utilisateur est pris en compte, le mode d'évaluation de la pertinence de l'information – uniquement technique, étude d'usages, *etc.*

1.2.4. Pour conclure sur « les contextes »

La question du contexte ne constitue pas le cœur de notre recherche, pourtant cette notion est omniprésente dans la littérature et le terme décrit souvent de nombreuses réalités. Pour Patrick Charaudeau et Dominique Maingueneau, « *le contexte d'un élément X quelconque, c'est en principe tout ce qui entoure cet élément* » (Charaudeau et Maingueneau, 2002 : p. 134). Si nous reprenons les trois cadres d'analyse précédents pour appréhender la recherche d'information, X peut être une unité linguistique, un dispositif technique ou une activité sociale.

- Lorsque X est une unité linguistique, l'entourage peut être l'environnement verbal de l'unité (appelé aussi cotexte) ou la situation de communication. (*ibid.)*. Nous parlerons alors de *contexte linguistique* pour désigner le cotexte et de *contexte discursif* pour désigner la situation de communication.
- Lorsque X est un dispositif technique, ce dernier peut être appréhendé de manière étroite par l'un de ces composants (par exemple l'interface) ou large (l'association de contenus, de technique et d'usages). Suivant les cas, l'entourage peut être interne ou externe au dispositif : l'utilisateur, la tâche, l'environnement social, le temps et l'espace, etc. sont autant de dimensions

134 La pagination est celle du document en ligne.

qui qualifient le contexte. Nous parlerons dans ces cas de *contexte social* ou *spatio-temporel, de contexte d'utilisation*, etc.

- Lorsque X est une activité sociale, l'environnement, perçu de manière subjective ou objective, est appelée *contexte* ou *situation*. Ainsi que le rappelle Céline Paganelli ces deux notions font référence dans la littérature à la recherche d'information, aux comportements de recherche des usagers ou encore à l'interaction entre les usagers et les systèmes (Paganelli, 2012 : 129). L'auteur relève cependant une différence d'emploi entre les deux termes :

> *« Il semble, toutefois, que la notion de* situation *soit davantage liée aux acteurs et à leurs buts, leurs activités, ou leurs habilités, quand le terme* contexte *est, en général, entendu comme faisant référence à un environnement plus large, constitué de facteurs et de variables qui affectent le processus d'information aux comportements de recherche des usagers ou encore à l'interaction entre les usagers et les systèmes. » (Paganelli, 2012 : 129)*

Par conséquent, nous réservons le terme de *contexte situationnel* aux éléments du contexte qui ont une influence sur l'activité informationnelle des individus, et utilisons comme précédemment le terme de *contexte social, contexte spatio-temporel* pour qualifier les autres dimensions qui affectent le processus d'information.

2. Des méthodes pour organiser l'information : le rôle des discours

Ce chapitre est consacré au déploiement de méthodes couramment utilisées en SIC pour recueillir et analyser les données. Elles sont issues de disciplines différentes comme les sciences sociales, les sciences cognitives ainsi que les sciences du langage, ou de techniques particulières utilisées par les SHS, comme l'analyse de contenu ou la lexicométrie. Nous montrons comment ces méthodes offrent une contribution à la recherche d'information, plus particulièrement, comment elles permettent d'analyser et de structurer des plans d'organisation de l'information. Ces méthodes font toutes appel à un matériau discursif, qu'il soit issu d'entretiens semi-directifs, de protocoles verbaux, de traces d'activités sur des documents. Ces discours sont tous enregistrés ou transcrits sur des supports. A ce titre, ils peuvent être constitués en corpus, ce qui est la condition indispensable au travail méthodologique. Tous les corpus recueillis

ont été nettoyés et documentés, cette démarche étant indispensable à la réplicabilité de l'analyse.

2.1. La liste des corpus analysés

Entre 2007 et 2013, nous avons recueilli et analysé 9 corpus. Dans la première colonne figurent, le cas échéant, les références des publications en lien avec l'analyse de ces corpus.

Réf.	Nom du corpus	Caractéristiques Dates de constitution	Constitué par	Analysé par
	Scientext-Mécanique	10 HDR et thèses de doctorat en mécanique recueillies sur support numérique pour enrichir le corpus Scientext. (recueil droit d'auteur) http://scientext.msh-alpes.fr/	Viviane Clavier en 2007[135]	Les membres de l'ANR Scientext
*ref. 31	Cancer-Presse	2446 articles de presse sélectionnés sur Europresse entre 2000 à 2005 et comportant les titres suivants : - Nouvel Obs Hebdo - Le Figaro - Le Monde	Hélène Romeyer en 2006	Hélène Romeyer et Viviane Clavier
	Cancer-Littérature	2706 extraits de textes sélectionnés dans Frantext entre le 16ème et le 20ème siècle relevant de plusieurs genres littéraires.	Viviane Clavier en 2007	Viviane Clavier
*ref. 35	Forum Santé	36 fils de discussion entre 2003 et 2007, extraits du forum de santé Doctissimo sur divers sujets en lien avec la santé. Corpus totalisant 879 pages, 6543 chaînes de caractères, 239 posts et 444 intervenants.	Viviane Clavier en 2008	Viviane Clavier, Céline Paganelli Evelyne Mounier, Adrian Staii
*ref. 39	Forum Santé maladies rares	12 fils de discussion entre 2005 et 2010 extraits du forum de santé Doctissimo, rubrique maladies rares et orphelines. Corpus totalisant 6311 chaînes de	Viviane Clavier en 2010	Viviane Clavier et Céline Paganelli

[135] Grâce à la contribution active d'Annie Leroy-Chesneau, enseignant-chercheur à l'Université d'Orléans, membre du Prisme.

		caractères, 415 posts et 222 intervenants.		
*ref. 37 40 41	Position-nement de l'auteur	153 passages de textes sélectionnés dans 10 thèses de doctorat et une HDR en sciences de l'information et de la communication	Viviane Clavier et Céline Paganelli en 2009	Viviane Clavier
ref. 34 et *ref. 38	Progrès Illustré[136]	389 causeries extraites du Progrès Illustré et 1475 illustrations entre 1890 et 1898, ainsi que 52 causeries scannées, numérisées et corrigées pour l'année 1900.	Viviane Clavier et Pierre-Yves Landron en 2010	Viviane Clavier
ref. 42 *ref. 46	Manuscrits de Stendhal	88 sites web comportant des manuscrits d'auteurs ou des informations sur les manuscrits	Viviane Clavier en 2012	Viviane Clavier
*ref. 44	Information profession-nelle	Corpus constitué de 2 glossaires (français, anglais), 3 terminologies (française, canadienne, européenne), 3 encyclopédies françaises, 206 articles de revues issus de Cairn.	Viviane Clavier en 2012	Viviane Clavier

La plupart des corpus que nous avons constitués ont systématiquement été exploités dans nos enseignements. Par exemple :

- en master 1 professionnel « métiers du livre » [137], nous avons utilisé le corpus « Cancer-Presse » pour aborder les phénomènes de polysémie des langues naturelles et plus largement l'indexation du sens figuré.
- en master 1 professionnel « communication scientifique et technique »[138], le corpus « Progrès illustré » a été mis à disposition pour réaliser des analyses de contenu et des analyses de discours sur la représentation des sciences et des techniques au 19ème siècle dans un journal populaire.

[136] http://collections.bm-lyon.fr/PER003100

[137] *« Traitement automatique des langues pour l'indexation »,* 15h TD, M1 IUP métiers du livre, Université Pierre Mendès-France (2004-2007).

[138] *« Méthodologie des SHS : analyse de discours et analyse de contenu* », 20h TD, M1 communication scientifique et technique, Université Stendhal (2008-10).

- en master 2 « sciences de l'information et de la communication »[139], nous avons exploité les introductions de thèses du corpus « Scientext », pour initier les étudiants aux méthodes de linguistique de corpus et d'annotation sémantique du lexique épistémique.
- en master 2 « recherche et études en information et communication »[140], Adrian Staii et moi-même initions les étudiants aux techniques d'annotation (TEI, Dublin Core) sur des corpus réels et volumineux (corpus « Progrès Illustré »).

2.2. Les entretiens et les protocoles verbaux

Hormis la constitution de corpus dont nous allons présenter les méthodes d'analyse ci-dessous, nous avons construit des grilles d'entretiens, passé, transcrit et analysé des entretiens. Deux études nous ont permis de nous forger une expérience dans ce domaine des sciences sociales.

Dans le projet collectif NOESIS, nous avons conduit 3 entretiens semi-directifs sur 16 [141] auprès de médecins spécialistes du CHU de Grenoble en juin et juillet 2006 afin de les questionner sur leurs usages des TIC et leurs pratiques de l'information médicale dans le cadre de leur activité professionnelle et scientifique (Balicco et al. 2007 ; Staii et al. 2008). Nous avons participé à l'exploitation de l'ensemble des entretiens. Dans le projet SCIENTEXT, nous avons menés 11 entretiens semi-directifs en binôme avec Céline Paganelli auprès de doctorants afin de les interroger sur leurs pratiques de recherche d'information et les usages qu'ils font des thèses de doctorat dans le cadre de leur activité de recherche. (Clavier et Paganelli, 2010). Par ailleurs, nous avons également été initiée à une méthode d'observation utilisant la technique des protocoles verbaux, issue de la psychologie cognitive (Bisseret et al., 1999) qui permet de « demander à des sujets de verbaliser, de penser tout haut » et de révéler les motivations et les procédures mises en œuvre par un sujet au cours d'une activité.

[139] *« Introduction au traitement automatique des langues* », 15h TD, M2 sciences de l'information et de la communication, option Information, usages et systèmes, Université Stendhal (2004-08).

[140] *« Gestion des connaissances »,* 12h CM (Viviane Clavier) et 8h CM (Adrian Staii), M2 recherche et études en information communication, Université Stendhal (2011-13).

[141] Nous étions 7 dans cette étude.

Cette observation a été réalisée à la suite des 11 entretiens mentionnés précédemment.

2.3. L'analyse des discours

Nous avons recouru à des méthodes et à des techniques pour analyser les discours recueillis dans des corpus « raisonnés », c'est-à-dire des corpus qui sont liés à un questionnement théorique. Nous évoquons ci-dessous la finalité des analyses conduites, la conception qui a présidé à la constitution de nos corpus et la manière d'aborder le contexte pour analyser les discours.

2.3.1. La finalité des analyses de corpus

Nos études sur corpus poursuivent deux objectifs. Soit elles présentent une réflexion de portée générale sur les méthodes et concepts utilisés en sciences de l'information et de la communication, soit elles offrent une contribution plus spécifique à la recherche d'information.

2.3.1.1. Des études de portée générale

Deux études relèvent du premier objectif : l'une porte sur la complémentarité des méthodes en information et en communication et montre le rôle du langage dans les discours médiatiques (Clavier et Romeyer, 2008). L'autre met à jour la terminologie en usage dans notre discipline (Clavier, 2013).

La première étude analyse les discours médiatique et littéraire sur le *cancer.* Cette recherche constitue un prolongement du travail de stage post-doctoral réalisé par Hélène Romeyer pour l'Institut National du Cancer (Romeyer, 2008)[142]. Dans l'article commun, nous avons envisagé la place des discours dans les méthodes d'analyse en SIC. Nous indiquions que l'information et la communication s'intéressent toutes les deux à la production et à la réception, des discours, mais qu'elles différaient sur les

[142] Romeyer Hélène (2008), *Les discours médiatiques du cancer en France (2000-2005) : étude menée dans le cadre d'un post doctorat à l'Institut National du Cancer,* Rapport de recherche, Institut National du Cancer-Département Sciences Humaines, Grenoble : Université Stendhal-Grenoble 3.

points d'entrée préalables à leur analyse : *« l'une [l'information] est centrée sur le langage et les effets de sens attribuables au code et l'autre [la communication] sur les effets de sens produits et reçus dans un espace donné. Dans les deux cas se pose la question du rapport entre le signifié de l'expression et l'interprétation du message. » (Clavier et Romeyer, 2008 : p. 2)*

Pour les sciences de l'information, l'analyse porte sur les textes, le langage et la recherche d'une signification stable, représentable et formalisable :

> *« Pour les sciences de l'information investies dans les traitements automatiques de l'information, la question du sens est abordée par la relation qui unit les formes de l'expression à leur signifié. Elle est travaillée dans le cadre des langages formels. Dans ce contexte, la signification d'un mot est mécaniquement déterminée par sa forme : à une phrase correspond un seul ensemble de conditions de vérité et à une expression linguistique, un seul ensemble de conditions d'applications. […] » (ibid.)*

Alors que pour les sciences de la communication, plus intéressées par les aspects énonciatifs et pragmatiques du discours, la focale porte sur les dimensions subjectives, sociales et historiques :

> *En sciences de la communication, ce n'est pas au signifié de l'expression que l'on s'intéresse, mais aux lieux et acteurs sociaux de sa production et de sa réception. Par conséquent, l'analyse n'entre pas par le texte lui-même, mais par le ou les discours entendus comme « trace qui a vocation à faire sens » (Utard, 2004) du point de vue des stratégies d'acteurs, des publics, des industries culturelles, dans des situations données. Ce qui intéresse ici c'est moins l'objet textuel en lui-même que l'analyse d'un ensemble de textes constituant une unité discursive. Ce n'est pas à des énoncés isolés que s'attache l'analyse mais à des ensembles de textes correspondant à une même situation d'énonciation : un corpus (Oger et Ollivier-Yanniv, 2007). » (ibid.)*

Pour montrer cette complémentarité, nous nous appuyions sur les conclusions d'Hélène Romeyer sur le discours médiatique sur le cancer « [qui] *utilise deux ressorts reliés directement aux peurs que suscitent la maladie et la mort : les risques et, son corollaire immédiat, la prévention* » (Clavier et Romeyer, 2008 : p. 4). Nous émettions alors l'hypothèse que la sémantique de la peur mise à jour dans la presse contemporaine apparaissait dans d'autres discours que les médias, qu'elle était ancienne, et qu'elle était ancrée dans la langue. Pour montrer cette inscription dans les discours sur le long terme, nous analysions les métaphores sur le *cancer* dans la

presse et la littérature et mettions en évidence le rôle du langage dans les représentations sociales sur un empan historique important, puisque les premiers textes littéraires remontent au 16ème siècle :

> *« Notre étude des discours sur le cancer révèle la nécessité de prendre en compte la part du langagier qui a un impact sur les représentations collectives de la maladie, travaillant sur le long terme de la mémoire discursive et sociale. Cette notion fait intervenir la dimension temporelle observée dans le processus de figement et la diversité des genres de discours. » (ibid.)*

La seconde étude s'intéresse à l'information professionnelle et fait l'objet d'un chapitre dans un ouvrage collectif co-dirigé avec Céline Paganelli sur ce sujet (Clavier et Paganelli, 2013). Ce chapitre a pour objectif de définir cette notion, de voir si elle est évoquée dans les discours des chercheurs et professionnels de notre discipline, et lorsqu'elle l'est, comment elle est abordée. Ce travail offre une contribution à la création « *d'outils intellectuels destinés à stabiliser le socle de références théoriques d'une discipline* » (Gardiès, 2011). Dans cette étude, nous recourons à des méthodes d'analyse de contenu et d'analyse de discours et mobilisons « *un appareillage théorique emprunté à la terminographie, comme les définitions, les concepts et les termes* » (Clavier, 2013 p. 49). Nous montrons dans l'étude de ces discours spécialisés que *l'information professionnelle*, contrairement à *l'information scientifique et technique*, et *l'information spécialisée* n'est ni un terme, ni un concept, ce qui fait de lui, « *un parent pauvre* » dans le champ de l'information.

2.3.1.2. Des études en recherche d'information

Hormis ces deux études, les autres offrent une contribution à la recherche d'information.

Tout d'abord, deux études sur le forum de santé Doctissimo abordent la question des forums comme sources d'information grand public (Clavier et al., 2010 ; Paganelli et Clavier, 2011). La première étude analyse les transformations que connait l'information de santé sous l'impulsion des techniques de communication numérique, passant du champ de l'information spécialisée à celui d'information grand public. Et la seconde, dans le prolongement de la précédente, analyse les modes de co-

construction de l'information au fil des échanges, identifie les formes d'hybridation d'informations de différents statuts : témoignages, vécu émotionnel, informations pratiques, informations médicales, etc. en adoptant les outils conceptuels et méthodologiques des sciences de l'information :

> *« [...] nous cherchons à mettre en évidence le caractère hybride de ces ressources en adoptant les outils conceptuels et méthodologiques des sciences de l'information. Ainsi, nous proposons d'analyser les formes d'entrelacs que revêtent les informations éditées et validées d'une part, et les échanges interpersonnels d'autre part. L'étude de ces effets de contamination est pour nous l'occasion d'une re-définition et d'une ré-interrogation de concepts clés en sciences de l'information : les propriétés de l'information, les notions de transfert ou d'usages de l'information que la forme même des forums fait évoluer. » (Paganelli et Clavier, 2011 : p. 40)*

Ensuite, une communication (Clavier 2008b) et trois publications écrites en collaboration avec Céline Paganelli (Clavier et Paganelli, 2010b ; Clavier et Paganelli, 2012a et b) envisagent la pertinence de la notion de positionnement de l'auteur pour la recherche d'information, cette notion étant également approchée en sciences du langage par nos collègues du projet SCIENTEXT. Le corpus, constitué de thèses de doctorat, permet de situer la place de ces documents dans les pratiques informationnelles des usagers, en l'occurrence des doctorants en SIC. Les méthodes d'entretiens et le recueil des traces d'activités sur les documents font le lien entre les pratiques informationnelles et l'indexation. Nos travaux contribuent ainsi à « *l'étude des pratiques informationnelles en contexte professionnel* » et ont pour objectif « *d'évaluer la pertinence de la notion de positionnement pour guider la consultation et l'annotation de documents scientifiques et à terme, pour en tenir compte dans l'indexation.* » (Clavier et Paganelli, 2010) Les publications suivantes offrent des précisions sur notre ancrage scientifique montrent plus précisément comment « *les connaissances linguistiques et cognitives en lien avec le positionnement peuvent être prises en compte pour améliorer l'accès à l'information* » (Clavier et Paganelli, 2012a et b), et formulent des « *propositions pour l'indexation et la représentation des connaissances véhiculées dans les discours scientifiques* » *(ibid.).*

Enfin, nous avons travaillé sur deux plates-formes d'archives du 19ème siècle, l'une concerne la presse illustrée (Projet CaNu XIX) et l'autre les manuscrits de Stendhal (Projet Métilde). Dans (Clavier, 2010), nous nous intéressons aux parcours

thématiques comme dispositif de mise en valeur d'un fonds patrimonial de presse ancienne :

> *« La création de parcours thématiques constitue l'un des moyens pour diversifier les modes d'accès aux collections, tout en répondant à des objectifs de valorisation et de mise en exposition du patrimoine de presse numérisée » (ibid., p. 114).*

La réflexion sur les parcours thématiques s'appuie sur l'analyse fine de 441 *Causeries*[143] numérisées et océrisées, une chronique fameuse du *Progrès Illustrée*, publiées entre les années 1890 et 1900.

Dans le projet des manuscrits de Stendhal qui est actuellement en cours de réalisation, nous participons avec d'autres collègues, à un programme destiné d'une part à identifier les publics spécialisés susceptibles d'utiliser la plate-forme, et d'autre part à envisager si cette plate-forme peut s'intégrer à leur pratique scientifique.

En conclusion, nos études offrent une contribution à la recherche d'information ainsi qu'aux méthodes et concepts mobilisées en SIC et font toutes appel à des analyses de discours, à l'exception de l'étude NOESIS.

2.3.2. Trois conceptions de corpus

Trois conceptions président à la constitution de nos corpus. Ils peuvent être constitués à partir de critères linguistiques, documentaires ou informationnels.

2.3.2.1. Des corpus linguistiques

Le corpus « Cancer » a été recueilli sur des critères sémasiologiques : tous les énoncés rassemblés dans ce corpus comportent une forme langagière : le mot *cancer*. Ce corpus comporte deux sous-ensembles qui ont été constitués en deux étapes, par deux personnes et analysés par des méthodes différentes.

[143] Le nom de la chronique du *Progrès Illustré*. La plupart des *Causeries* ont été rédigées par les éditorialistes Jacques Mauprat et Paul Clairfont.

Un premier corpus, que nous appelons « Cancer-presse », a été collecté par Hélène Romeyer en 2006 et rassemble 2446 articles issus du *Monde*, du *Figaro* et du *Nouvel Obs* entre 2000 et 2005, période marquée par la politique de Jacques Chirac sur la lutte contre le cancer (Romeyer, 2008). Méthodologiquement après avoir éliminé de son corpus les valeurs figurées du mot *cancer* (les métaphores) et les homonymes (le signe astrologique), l'auteur procédait à une analyse de contenu et mettait en évidence le lien entre des discours sur la maladie, la mort, les risques et la prévention. Ce corpus ne fait pas appel à des méthodes d'analyse linguistique.

Un second corpus de 2706 extraits de textes littéraires dénommé « Cancer-Littérature » a été constitué en 2007 par nos soins à partir de la base de données Frantext. Il fait appel à des méthodes linguistiques pour collecter et analyser les énoncés en diachronie. Il s'agit d'analyser les valeurs propres et figurées (métaphores, personnification) de l'unité lexicale simple *cancer* ou complexe *(cancer du sein, du colon),* et de ses formes dérivées *(cancéreux, cancérologie, oncologie*). L'étude consiste à analyser le contexte de ces occurrences et à déceler des régularités sémantiques. Les contextes gauche et droit présentent des items lexicaux qui font référence à la mort, la déchéance et la souffrance, que ce soit dans les discours littéraires et la presse, en diachronie et en synchronie. Le corpus littéraire mettait en évidence l'existence de « *métaphores usuelles, figées, voire mortes* » (*ibid.* : p.3) sur le *cancer* qui traversent le temps jusqu'à notre époque.

Nous voulions alors suggérer l'existence d'un continuum et d'un lien, entre d'un côté les discours (littérature, presse) et de l'autre, la formation d'une mémoire sociale qui a « *un impact sur les représentations collectives de la maladie* » (*ibid.)* :

> *« Cette pathologie dispose d'un fort capital symbolique dans la mémoire sociale : elle est associée à la mort, la souffrance, à une lutte inégale, à quelque chose qui 'ronge de l'intérieur'. Ces représentations négatives sont transmises au mot cancer lorsque celui-ci ne dénomme pas la maladie mais d'autres fléaux : le cancer du chômage, le cancer terroriste » (ibid. : p.1)*

Pour adhérer à ces propositions, il faut faire l'hypothèse que les discours littéraires peuvent être traités avec les mêmes outils que les autres discours sociaux, ce qui est loin d'être une évidence ainsi que le note Dominique Maingueneau. En effet, les

premiers sont des événements esthétiques, les seconds des techniques d'écriture ordinaire :

> *« Comme le dit de son côté A. Herschberg Pierrot, « le discours est opposable à l'œuvre littéraire. L'œuvre n'est pas un discours parmi d'autres, c'est un événement d'écriture et de lecture et une configuration esthétique […] Dans cette perspective, l'analyse du discours et celle du style n'ont pas les mêmes enjeux ni ne portent sur les mêmes objets. » Ce qui conforte évidemment une certaine distribution des tâches dans l'univers académique : la distinction entre les facultés de lettres, qui auraient en charge les œuvres, et les sciences humaines et sociales, naturellement portées à l'étude des textes de second plan. (Maingueneau, 2008).*

En mettant sur le même plan littérature et presse, nous ne considérons pas que ce sont des « discours » comparables mais nous montrons comment la littérature pourtant si créatrice dans ces productions, contribue pourtant à installer des représentations durables qui finissent par s'inscrire dans la langue. Ainsi, l'analyse des cooccurrences met à jour des routines discursives dans l'utilisation de nombreuses métaphores figées, alors que d'ordinaire, la métaphore est un lieu de créativité littéraire. Ces métaphores usées finissent par être « absorbées » dans le langage sous forme de dénominations :

> *Durant ce processus graduel de lexicalisation qui conduit au figement et à son absorption dans la langue, la mémoire semble jouer un rôle prépondérant. Pour Perrin (2002) en effet, « tout discours regorge de formes linguistiques assorties de significations référentielles préalablement mémorisées, héritées de faits pragmatiques ponctuels, de figures rhétoriques la plupart du temps, souvent métaphoriques, dont on a pris l'habitude et que l'on finit parfois par entériner sous forme de dénominations. » (Clavier et Romeyer, 2008 : p. 3)*

L'étude des valeurs métaphoriques de *cancer* révélait alors l'ancienneté et la permanence des représentations sociales de cette maladie dans les discours : la peur, la mort, la souffrance, la dégénérescence. Ces informations pouvaient utilement être prises en considération dans les campagnes de prévention.

Bien que nous ayons recouru à une méthode classique en linguistique inspirée du distributionnalisme harrissien, notre approche n'est pas vraiment linguistique, puisque nous négligeons l'impact de la forme langagière sur le sens pour ne retenir que la signification, ce qui, pour un linguiste serait irrecevable. En revanche, cette approche relève d'une conception informationnelle du sens.

2.3.2.2. Des corpus documentaires

Les textes et énoncés – essentiellement des définitions – rassemblés dans les corpus « Information professionnelle », « Forum santé » et « Forum santé maladies rares » ont été constitués à partir de critères documentaires. Nous signifions par là, qu'ils proviennent de sources documentaires utilisées par un public spécialisé ou par le grand public. Nos méthodes d'analyse participent de procédures mises en œuvre en documentation pour évaluer et caractériser les ressources : le statut de l'information est pris en compte, ainsi que celui de l'auteur, les dates, l'organisation du contenu, les informations sont croisées et mises en perspective…

Les textes du corpus « Information professionnelle » sont issus d'ouvrages de référence contemporains spécialisés. Des glossaires, des dictionnaires encyclopédiques spécialisés en information-documentation, des terminologies, des articles de revues spécialisées en SIC et économie et gestion ont été recueillis. Ces sources résultent de la production scientifique d'une communauté d'acteurs professionnels et scientifiques en information-documentation ou en SIC. A ce titre, les discours analysés sont révélateurs de pratiques socio-discursives comparables (du moins institutionnellement) et sont un lieu d'observation des emplois de *l'information professionnelle* comme un terme, une entrée vedette, une définition ou un concept. Voici quelques exemples illustrant les dimensions prises en compte pour analyser ce corpus.

Un premier exemple réside dans la prise en compte des traditions éditoriales et scientifiques ayant présidé à la confection de ces ressources. Nous considérons en effet, qu'il est nécessaire de préciser le statut des documents consultés pour appréhender des notions. Par exemple, à propos des encyclopédies et des terminologies :

> *« Pour cerner des notions, donner des définitions, préciser le sens de vocabulaires, on peut utiliser des terminologies ou des encyclopédies, appelés aussi dictionnaires encyclopédiques. Terminologies et encyclopédies correspondent à des traditions différentes. La confection de terminologies a évolué d'une conception classique, à laquelle se rattache l'approche*

> *documentaire, vers une conception plus linguistique prenant en compte les discours. » (Clavier, 2013 : p. 50)*

La confection de ressources terminologiques a été directement influencée par des méthodes scientifiques impliquant les discours. Par exemple, les terminologies ont évolué « *d'une conception classique, à laquelle se rattache l'approche documentaire, vers une conception plus linguistique prenant en compte les discours.* » (*ibid*. p. 50). Quant à la tradition encyclopédique, nous précisions que « *l'encyclopédie ne fournit pas de définitions à proprement parler, mais plutôt des bilans de connaissances* » (*ibid.)* et qu'elles sont des lieux « *d'enregistrement, de consécration des rapports de force intellectuels, institutionnels, voire un lieu d'innovation [DUM 94]* ». Nous évoquions également la difficulté à capter le « *socle de références communes* » au sein des sources de références tant la conception qui préside à la délimitation des vocabulaires est floue dans les sources de référence :

> *« Les glossaires, les terminologies, les dictionnaires encyclopédiques permettent d'appréhender le vocabulaire d'un métier, d'un secteur professionnel ou d'une discipline, soit le vocabulaire d'un domaine. Ces ressources poursuivent le même objectif : produire des listes alphabétiques […] ainsi que des définitions. » (ibid., p. 51-52).*

… et que les connaissances y sont rapidement périmées :

> *« Toutes ces ressources ont l'inconvénient de devenir d'autant plus vite obsolètes que le domaine à décrire est actif, la productivité lexicale étant liée aux pratiques sociales et aux discours. » (ibid., p. 52)*

Un second exemple s'attache à la prise en compte de l'inscription disciplinaire des auteurs des articles de revues (information-communication ou économie et gestion), de leur statut (professionnel, chercheur), de la perspective dans laquelle est écrit l'article (article de fond, étude de cas, présentation de méthodes et techniques, compte rendu d'ouvrages), de la problématique envisagée, du cadre de la recherche, etc. Cette perspective a conduit à une grille d'analyse fouillée pour éclairer l'interprétation de l'unité lexicale « *information professionnelle* » en tenant compte de l'ensemble de ces dimensions.

Les textes des corpus, « Forum santé » et « Forum santé maladies rares », sont des fils de discussion extraits de Doctissimo. Les forums de discussion sont prisés par le

grand public. En 2010, l'observatoire des usages internet de Médiamétrie décomptait *« plus de 13 millions d'internautes [qui avaient] lu des messages sur des forums, soit 10% de plus que l'année précédente. »* (Paganelli et Clavier, 2011 : p. 39). Cependant, en comparant le nombre de messages postés et le nombre de consultations sans intervention explicite sous forme de questions, nous constations que les forums pouvaient aussi être utilisés comme source d'information et non exclusivement comme outil de communication :

> *« En effet, alors que le nombre d'interventions est relativement faible (409) le nombre de fois où les fils sont consultés sans intervention en revanche, est très important (130 536), soit 319 fois plus que le nombre d'interventions. Dès lors on peut supposer que des individus consultent les forums pour trouver une information déjà traitée dans les messages précédents sans avoir de question précise à poser.» (ibid., p. 44).*

Nous observions que ces sources d'information étaient alors considérées comme un réservoir d'archives alors que la plupart des études réalisées sur ces dispositifs « *s'attachent à l'étude des échanges asynchrones […] sous l'angle de l'analyse conversationnelle et interactionnelle*» (Clavier et *al*., 2010 p. 298). Nous avons proposé de nouvelles méthodologies d'analyse inspirées des SIC. Ces méthodes ont consisté à typer les échanges d'information[144] mettant en évidence « de nouvelles dynamiques interactionnelles » spécifiques à ce type de source, comme l'enchaînement polyphonique de questions et réponses, la présence de différentes catégories d'information : information scientifique, témoignage, interpellation, information pratique, conseils. Ensuite, ces ressources ont été analysées sous l'angle terminologique afin d'évaluer la fiabilité des informations[145] : « *une terminologie rigoureuse étant révélatrice d'une information médicale fiable* » (ibid. p. 305). En comparant le lexique médical au Medical Subject Headings (MeSH), nous avons analysé le vocabulaire spécialisé et ses variantes. Enfin, constatant que les témoignages constituaient le type d'information le plus représenté, une étude des formes que prennent ces témoignages a été réalisée[146]. Elle met en évidence deux types de témoignages : les récits de vie ainsi que les descriptions d'états psychologiques et pathologiques…

[144] Par Evelyne Mounier, Céline Paganelli et Adrian Staii.

[145] Par nos soins.

[146] Par Maria-Caterina Manes-Gallo.

Dans (Paganelli et Clavier, 2011), la même méthodologie a été appliquée sur un corpus de fils plus spécialisés afin de « *resserrer l'étude sur une famille de pathologies pour lesquelles les sources d'information sont plus rares et moins vulgarisées* ». (Paganelli et Clavier, 2011 : p. 42). Nous mettions en évidence le caractère hybride de ces ressources qui présentent des points communs avec les blogs de santé : ils sont un lieu d'échanges intermédiaire entre le cabinet médical et le groupe de parole. Un typage plus exhaustif des niveaux d'information est présenté ainsi que « *les formes d'entrelacs que revêtent les informations éditées et validées d'une part, et les échanges interpersonnels d'autre part.* » (*ibid.* p. 40).

Les analyses de ces forums révélaient la présence de nombreuses marques d'émotion :

> *« Nous appelons lexique émotionnel une catégorie sémantique qui renvoie essentiellement à des processus psychologiques (Grossmann et Tutin., 2005). On relève par exemple des adjectifs évaluatifs (cloques énormes, petites nausées), des comparaisons qui indiquent le caractère hors norme d'une situation (cloques grosses comme des balles de ping-pong). L'apparition de cette catégorie de lexique est le signe d'une intrication des dimensions médicales et émotionnelles […] » (Clavier et al. 2010 : p. 307).*

Se manifestant sous diverses formes langagières, ainsi que sous la forme d'icônes (les smileys) et de ponctuations, les marques d'émotion se mêlaient aux informations médicales. Il nous semblait que cette forme « d'entrelacs » participait pleinement du succès des forums de santé et contribuait aussi à redéfinir les contours de la recherche d'information dans le domaine de la santé. Ce résultat a fait l'objet du sujet de la thèse de doctorat de Céline Battaïa.

2.3.2.3. Des corpus « informationnels »

Enfin, nous avons mis en œuvre une troisième catégorie de corpus destiné à recueillir des discours : les corpus « informationnels ». Nous signifions par cette dénomination que les corpus sont recueillis *via* un processus expérimental mis en place pour observer les usages effectifs de documents (ou de dispositifs informationnels) lors de la recherche d'information. Un « corpus informationnel » rassemble des données

collectées suivant diverses méthodologies qui sont enregistrées sur un support. Les données proviennent de trois sources : les documents, les traces d'activité laissées par les lecteurs lors de la consultation de documents ainsi que les commentaires des lecteurs sur leurs pratiques.

Cette conception a été mise en œuvre dans le corpus « Positionnement de l'auteur ». Les documents concernés sont les thèses de doctorat en SIC et les lecteurs sont des doctorants également en SIC, qui consultent ces documents pour écrire leur propre thèse. La première publication écrite en 2010 décrit précisément la méthode de recueil des corpus (Clavier et Paganelli, 2010)[147] :

> *« La méthodologie mise en place est double. Dans un premier temps, nous recourons à des entretiens et observations de sujets en situation de consultation de thèse, en utilisant la technique des protocoles verbaux, issue de la psychologie cognitive (Bisseret, 1999). Cette technique qui consiste à demander au sujet de verbaliser, penser tout haut, permet de révéler les procédures utilisées par le sujet au cours de son activité. Les sujets sont des doctorants et enseignants-chercheurs. Les résultats permettent de répondre à diverses questions […] et les observations de recueillir un corpus de fragments de textes extraits des thèses et jugés pertinents par les sujets. Dans un second temps, nous recourons à un typage des fragments, mentionnons la taille, le nombre et leur position dans la structure du document ; nous décrivons le statut du fragment en lien avec la structure du document ou des unités de langue ; enfin, nous nous concentrons sur les marqueurs de positionnement » (ibid.)*

Ainsi, nous nous sommes intéressées aux documents consultés par 11 usagers en situation de recherche d'information que nous avons interrogés sur leur lieu de travail. Dans les documents consultés, seules les informations jugées pertinentes par les lecteurs ont été retenues. La plupart du temps, les lecteurs travaillaient sur les thèses imprimées ou photocopiées et non sur les supports numériques à l'écran ; les passages intéressants étaient surlignés, les notes étaient prises dans un document différent (fichier, carnet de notes, post-it) ou sur la thèse elle-même. Ce sont ces traces écrites qui nous révélaient la manière dont les usagers s'approprient individuellement le document :

[147] Toutefois, la notion de « corpus informationnel » n'a pas été formalisée ni présentée comme telle dans nos publications.

> *« Aux passages de texte sélectionnés, les lecteurs rajoutent souvent des annotations : commentaires, mots-clés, références bibliographiques en lien avec le thème. Certains sujets mettent en oeuvre un système de codification assez précis. Les annotations représentent ici la perception que l'annotateur a du document qu'il consulte ; elles sont une manière pour lui de s'approprier le document et d'en interpréter le contenu (Mille, 2005). C'est bien une lecture propre à chaque individu qui est menée. » (ibid.)*

En outre, nous nous sommes intéressées aux discours oraux recueillis dans le cadre des entretiens et des observations. Ces discours ont été transcrits et éclairent de manière plus globale les stratégies de consultation des thèses de doctorat par les sujets, ou de manière plus ciblée, le choix de tel ou tel passage de texte sélectionné, son intérêt pour la tâche à réaliser, son rôle dans la construction d'un raisonnement scientifique, etc. :

> *« L'observation des parcours de consultation des thèses de doctorat a mis en évidence une lecture fragmentée, où les sujets lisent des extraits choisis en fonction de la tâche qu'ils réalisent, de l'état d'avancement de leur propre travail. La consultation de ces documents s'inscrit donc pleinement dans le cadre de la lecture professionnelle. Seule la lecture apprentissage observée chez les sujets de première année et la lecture disciple observée lorsqu'une thèse fait unanimement référence dans le champ disciplinaire, échappent à ce constat. » (ibid.)*

L'analyse précise du document annoté indiquait que la lecture était accompagnée d'annotations uniquement lorsqu'il y a des enjeux de positionnement et que l'annotation semble favoriser la compréhension et la construction d'une identité scientifique singulière.

En conclusion, la notion de corpus « informationnel » mérite une plus grande conceptualisation que la dénomination actuelle, sans doute maladroite, ne peut clarifier. Par cette dénomination, nous voulons mettre en avant le recours à une méthode qualitative, fondée sur un travail de terrain, avec une volonté affichée d'interprétation, ce qui nous semble relever de principes ethnographiques. Par ailleurs, nous appréhendons des objets dans le cadre de pratiques sociales clairement identifiées, ce qui relève de principes sociologiques. Cependant, l'expérimentation est menée en laboratoire, ce qui se rapproche des méthodes expérimentales conduites en psychologie cognitive… Quoi qu'il en soit, l'intérêt d'un corpus « informationnel » se situe à deux niveaux. D'une part, il permet le recueil de diverses « sources de

traçages », pour reprendre l'expression de (Yahiaoui et al., 2008). La prise en compte des discours sur les pratiques informationnelles autorise leur « redocumentation » qui permet non seulement d'expliquer le contexte de réalisation des activités informationnelles, mais de donner un sens aux traces. D'autre part, un corpus « informationnel » introduit plus nettement dans le cadre d'observation des pratiques, les objets de la recherche. **Dans le cadre de l'étude réalisée, c'est le document comme complexe d'observation et lieu d'inscription de connaissances qui est mis au premier plan.** En revanche, cette méthodologie souffre de limites. Outre son caractère micro-social, cette méthodologie n'est pas toujours facile à mettre en œuvre, certains sujets ne souhaitant pas toujours « montrer leurs notes ». Elle nécessite un travail d'annotation complexe superposant divers niveaux de traces issues de sources différentes et produites à des moments distincts : pendant et après la consultation du document, au moment des entretiens. A cet égard, notre système de typage des fragments éclairé par l'ensemble de ces traces est trop artisanal : nous avons utilisé des champs de description dans des fichiers Excel sans recourir aux outils de la TEI qui auraient permis à d'autres chercheurs d'exploiter le corpus. Enfin, cette conception des corpus est éphémère et dénuée de valeur intrinsèque, contrairement aux corpus oraux transcrits par exemple ou aux corpus documentaires présentés *supra*.

2.3.3. Analyser le contexte : un construit méthodologique

Nos études présentent le point commun de faire appel au contexte, notion que nous avons introduite précédemment (cf. 1.2.) Nous appréhendons le contexte comme un construit méthodologique destiné à analyser des discours. Deux conceptions du contexte ont été mises en oeuvre. La première prend comme point de départ des formes langagières dont nous analysons l'environnement. La seconde crée des situations d'observation « en laboratoire » qui réunissent des éléments convoqués dans une activité de recherche d'information. Dans les deux cas, le contexte apparaît alors comme un construit méthodologique : c'est nous qui déterminons le statut des composantes contextuelles, qui en définissons la portée et la fonction.

2.3.3.1. Appréhender une forme langagière dans son environnement linguistique *versus* discursif

Lorsque notre point d'entrée est une forme langagière, dont nous voulons appréhender les usages en discours, le contexte se résume à l'environnement étroit (le co-texte) et consiste à mettre à jour des cooccurrences lexico-sémantiques d'un terme-pivot. Ainsi, pour analyser le corpus « Cancer », nous avons recouru à des méthodes distributionnelles où seul le matériau langagier est analysé. Nous avons sélectionné les environnements gauche et droit de *cancer*, le terme-pivot, en limitant la taille du contexte à la phrase. Nous avons recueilli les éléments lexicaux qui se trouvent fréquemment associés dans le voisinage du terme-pivot et avons procédé à l'interprétation des distributions obtenues. Nous avons ainsi constitué des classes d'équivalence par types d'expansions de *cancer* afin d'identifier les organes les plus souvent évoqués : après le *cancer du sein,* les organes se diversifient au 18ème siècle : *le sein, le foie, la gorge, la lèvre, l'estomac, le pancréas*. Puis, nous avons également recueilli des classes d'adjectifs évaluatifs (*cancer infâme, monstrueux, sérieux, affreux*), des classes d'adjectifs épistémiques (*cancer implacable, incurable, inévitable*) ou des classes de verbes imperfectifs (*ronger, pourrir, dévorer, généraliser, proliférer, métastaser, résister),* etc. Nous avons fait appel à des outils de lexicométrie pour recueillir des cooccurrences : le concordancier Antconc et le détecteur de cooccurrences lexicales Cooc [148]. Pour réaliser cette étude, les textes ont été lemmatisés avec le logiciel Cordial Synapse Développement, puis mis au format de Lexico 2[149] afin de pouvoir être traité par Coocs. Le traitement est un peu long, mais l'intérêt de Coocs est qu'il permet non seulement de déterminer la taille du contexte, mais de préciser le seuil de co-fréquence de l'unité du contexte à proximité du « mot-pôle », ainsi que de donner un seuil de spécificité : les résultats sont visualisables sous la forme d'un graphe (cf. Annexes 3 et 4).

Dans le corpus « Information professionnelle » en revanche, le contexte pris en compte est plus large et intègre des informations éclairant la production du document : la tradition éditoriale, le statut des auteurs, le laboratoire, les enjeux

[148] Antconc est développé par Laurence Anthony et Coocs par Williams Martinez.

[149] Développé par le Centre de Lexicométrie et d'Analyse Automatique des Textes (SYLED-CLA2T).

d'écriture, etc. Le *contexte discursif* comporte aussi des informations issues du paratexte : le résumé, les notes de bas de pages, les commentaires, les mots-clés, etc. Dans ce cas de figure, l'environnement est moins restrictif, il est propice à des analyses de contenu ou à des approches qui dépassent le cadre phrastique. Ainsi, pour comparer les « vocables » *information professionnelle*, *information scientifique et technique* et *information spécialisée*, nous avons distingué les différents niveaux de contextes et avons catégorisé ces discours, à la manière de l'analyse de contenu :

> *« La notion de contexte peut être réduite à l'environnement immédiat de l'occurrence à analyser ou faire référence à des éléments de connaissance éclairant la production du document lui-même. La première approche est développée par l'analyse de discours. Elle permet de travailler les significations du vocabulaire en exploitant le contexte (le plus souvent on recourt à des statistiques lexicales comme la méthode des mots associés). La seconde approche est adoptée par l'analyse de contenu, elle permet de travailler les concepts indépendamment des usages linguistiques. Nous allons tenter de combiner ces deux approches. (Clavier, 2013 : p. 60)*

A l'issue d'un travail d'interprétation des contextes discursifs associés aux trois occurrences, nous avons obtenu 7 catégories dotées d'une dénomination et d'une définition, chaque occurrence de vocables relevant d'une seule interprétation étiquetée au moyen d'une catégorie (Clavier, 2013) :

- ORGANISATIONS : l'information est citée en lien avec des organismes, agences, entreprises ;
- CONNAISSANCES : l'information est abordée sous l'angle de connaissances traitées, mises en forme ;
- CONTENU : l'information est envisagée sous l'angle de données, de langages ;
- PUBLICISATION ET RESSOURCES : l'information renvoie à des types de sources et fait l'objet d'une « publicisation » ;
- METADISCOURS SCIENTIFIQUE : l'information fait l'objet de discussions, de définitions, de conceptualisations ;
- MARCHE : l'information est vue sous l'angle d'un marché, d'une filière industrielle ;
- PRATIQUES PROFESSIONNELLES : l'information est évoquée sous l'angle des activités et pratiques professionnelles.

Cette méthode a permis d'apprécier la diversité des contenus associée aux vocables et de comparer les vocables entre eux. *L'information spécialisée* couvre 6 catégories sur 7, alors que *l'information professionnelle* 4 sur 7, ce qui laissait supposer que le premier terme était le plus général. Ce principe de catégorisation caractéristique de

l'analyse de contenu *« gomme la diversité des acceptions que l'on peut rencontrer dans les textes* » (*ibid*., p. 63). Nous avons alors pris en compte la fréquence d'apparition des vocables dans les discours, observant ainsi que *l'information spécialisée* était l'expression la moins usitée dans les discours spécialisés, alors que son usage est recommandé. Nous avons également identifié pour chaque vocable employé, le statut de l'auteur de l'article, dans quelle discipline il évolue, dans quel type de problématique il s'inscrit, les thématiques abordées, etc. Cet examen montre que seule *l'information scientifique et technique* jouit d'une forte visibilité, qu'elle est souvent évoquée en lien les politiques publiques de la formation et de la recherche, qu'elle présente un ancrage historique fort et enfin, que l'INIST est l'acteur de référence, ce dernier ayant joué un rôle central dans la diffusion et le traitementde l'IST. Grâce à la prise en compte de ces éléments de connaissance, nous avons été en mesure de donner un sens aux fréquences d'emploi de cette occurrence, ce qui n'aurait pas été possible avec des méthodes uniquement statistiques.

2.3.3.2. Appréhender un document dans le cadre de pratiques informationnelles

Dans l'étude sur le positionnement scientifique, nous avons créé une situation d'observation qui contextualise l'usage d'une catégorie de documents, les thèses de doctorat. Ces documents présentant la particularité de révéler la posture scientifique de chercheurs en formation. Dès lors, ce sont les usagers eux-mêmes qui commentent les passages qu'ils ont sélectionnés dans un document, les raisons de cette sélection, ce qu'ils font de l'information retenue, etc. La méthode de constitution des « corpus informationnels » détaille la manière dont nous recueillons des traces issues de différentes sources. Seules les publications sur la consultation des thèses de doctorat mettent en œuvre cette approche, elles présentent la caractéristique d'accorder au document un rôle central pour faire le lien entre les pratiques et les traitements.

Les documents jouent alors un double rôle : soit, ils sont l'une des composantes du contexte d'observation des pratiques informationnelles. Soit, ils sont l'objet de l'indexation. Dans la répartition des tâches que Céline Paganelli et moi-même nous attribuons, ma collègue s'intéresse plutôt au premier aspect, et moi au second.

Ainsi, l'intérêt de Céline Paganelli « *se focalise sur la manière dont [les] usagers, dans leur contexte professionnel, interagissent avec l'information, qu'il s'agisse de production, de traitement, de recherche, de partage, et ce en lien avec le contexte organisationnel* » (Paganelli, 2012 : p. 64). Elle s'intéresse aux activités informationnelles qui sont « *des activités sociales influencées par divers facteurs inhérents au contexte et à la situation dans laquelle elles prennent place* » (Paganelli, 2013 : p. 223). Ces activités sont saisies dans le cadre de l'observation de pratiques informationnelles, ainsi que diverses dimensions situationnelles dont l'auteur souligne l'importance :

> *« Il nous semble toutefois important d'affirmer clairement les dimensions sociales, temporelles et spatiales des pratiques, dimensions qui sont souvent absentes des recherches consacrées à l'accès à l'information. » (Paganelli, 2012 : p. 42-43)*

Plus précisément, les facteurs qui jouent un rôle important sur la recherche d'information sont l'activité principale de l'usager, l'organisation dans laquelle prennent place ces activités ainsi que l'environnement social qui intègre les pratiques culturelles plus larges. Ces éléments ont été mentionnés dans nos articles communs à plusieurs reprises et ont des conséquences sur les stratégies de lecture des documents.

> *« Depuis nos premiers travaux (Clavier, 1997), nous cherchons à combiner deux approches de l'activité d'information : nous nous fondons sur les usages et sur les caractéristiques textuelles des documents pour définir les traitements. Cette perspective, que l'on peut qualifier de recherche d'information « contextualisée » est partagée par d'autres chercheurs même si les paramètres de contextualisation sont variables d'un auteur à l'autre. En ce qui nous concerne, nous étudions les pratiques de recherche d'information de manière située, considérant que l'activité principale de l'individu influence les attentes en matière d'information, les stratégies de recherche et plus largement l'activité informationnelle » (Clavier et Paganelli, 2012 : p. 172)*

En ce qui nous concerne, nous recueillons les informations utiles pour guider l'indexation et tenons compte des traces d'activités sur les documents, éclairés par les pratiques informationnelles :

> *« Nous considérons ainsi que la conception d'un système nécessite que soient prises en compte les spécificités du contexte de la tâche de recherche d'information (Paganelli et al, 2002), ainsi*

> *que les caractéristiques textuelles et discursives des documents (Poudat et al, 2006) » (Clavier et Paganelli, 2012 : p. 172)*

Plusieurs études envisagent une indexation automatique du texte intégral : nous avons situé nos travaux comme une étape préalable à l'indexation :

> *« L'analyse des pratiques et l'étude des propriétés linguistiques et structurelles des documents apparaissent alors comme des préalables nécessaires à l'indexation. » (Clavier et Paganelli, 2012 : p. 173).*

Ainsi, dans l'étude conduite sur les thèses de doctorat, nous avons caractérisé formellement et structurellement les passages sélectionnés par les usagers lors de la consultation des documents. Le but poursuivi est un typage des données langagières en regard de critères déterminants pour l'indexation automatique tels que définis dans la première partie de ce document. Cette analyse nous a conduit à caractériser les données suivant différents plans d'organisation de l'information, comme les niveaux du discours et du texte :

> *« Nous avons analysé 158 fragments de textes. 129 d'entre eux comportaient des marques visuelles (soulignement, surlignage, etc.), 47 présentaient des annotations (notes, abréviations, mots-clés, symboles ; 148 ont été commentés oralement. Nous avons observé que les annotations et les commentaires permettaient d'identifier deux ensembles d'indices dans les fragments. Le premier ensemble opère au niveau du discours, l'autre au niveau de la textualité. » (Clavier et Paganelli, 2012 : p. 174)*

En outre, nous avons pris en compte la position des fragments au sein de la structure du document. Parmi les résultats, nous observions que la taille moyenne des fragments est de 100 mots, qu'ils se répartissent à tous les niveaux de structure du document, y compris les notes de bas de page, qu'il n'y a pas de lien entre la taille des fragments et les tâches, et que la taille des fragments ne varie pas avec l'ancienneté dans l'activité de recherche. Nous remarquions également que la quasi-totalité des fragments correspond à un niveau de structure du document, et que le niveau le plus répandu était le paragraphe. Nous notions cependant, que les paragraphes retenus n'étaient pas toujours complets.

2.3.4. Conclusion sur la place de la méthodologie dans notre recherche

En conclusion, la place que nous accordons aux méthodes est toujours très importante, parfois sans doute pesante, alors qu'une grande partie de la production scientifique en SIC semble au contraire accorder plus de place aux études théoriques. C'est ainsi que Bernard Miège (repris par Simone Bonnafous et François Jost) juge les travaux conduits dans notre discipline qu'il taxe de «*productions théoriques à visée généraliste se proposant de couvrir l'ensemble du 'champ' de la communication (sociale ou microsociale)* »[150]. Il convient cependant de préciser quels contours donner à la notion de méthodologie : pour nous, relève de la méthodologie, l'agencement des méthodes en regard de l'interdisciplinarité revendiquée par les SIC (Ollivier, 2001), ce que nous indiquions dans la première publication réalisée au GRESEC :

> *« Finalement, en SIC, étudier les discours, c'est donc se situer à l'articulation d'une pratique discursive et de pratiques sociales. L'analyse des discours suppose ainsi à la fois un travail sur l'énonciation (son contexte et son auteur), une démarche d'analyse linguistique des énoncés, une démarche d'analyse sociologique des énoncés, une étude des stratégies d'acteurs et des logiques sociales à l'oeuvre. « Le rapprochement de la sociologie et de l'analyse de discours est particulièrement pertinent pour l'étude des phénomènes de communication » Miège (2000 : 557- 559) » (Clavier et Romeyer, 2008 : p. 2)*

En premier lieu, nous avons réalisé de nombreuses études empiriques dans différents terrains, mêlant des méthodologies « *reconnues par les sciences humaines et sociales* » (Ollivier, 2007 : p. 169). Bruno Ollivier rappelle qu'elles sont au nombre de quatre : l'étude des documents, l'observation, l'enquête quantitative et l'entretien ». Nous en avons utilisé trois sur quatre[151].

En second lieu, à l'exception des deux études sur le cancer et sur l'information professionnelle, nous avons abordé les méthodes en lien avec la recherche d'information. Trois ensembles se dégagent :

[150] Extrait cité par Simone Bonnafous et François Joste qui sont du même avis que Bernard Miège pour dénoncer cette posture généralisante qui « *ne se donne pas les moyens de valider les propositions formulées* » (Bonnafous et Jost, 2000 : p. 537),

[151] Nous n'avons jamais réalisé d'enquête quantitative.

- Un premier ensemble de méthodes porte sur l'observation de pratiques informationnelles d'individus en contexte de travail. Ce sont les études conduites en collaboration dans le cadre de recherches finalisées : les projets NOESIS, SCIENTEXT, Métilde. Les entretiens se sont déroulés auprès de spécialistes confirmés en médecine et en littérature, ou auprès d'apprentis-chercheurs comme les doctorants en SIC. Ces études offrent une contribution à l'évaluation de dispositifs existants (Clélia pour les manuscrits de Stendhal), en cours de développement (NOESIS au CHU de Grenoble), ou comparables (Thèses en Ligne, Cyberthèses), ainsi qu'une contribution théorique sur les pratiques informationnelles en contexte professionnel.
- Un second ensemble de méthodes porte sur des dispositifs (les forums de santé), ou des documents (encyclopédies, glossaires, terminologies, articles de revues) et sont appréhendés comme ressources documentaires. L'enjeu méthodologique porte sur la mise en évidence de différents plans de structuration de l'information et sur la prise en compte du statut éditorial de ces ressources. **L'information est alors appréhendée comme « *la relation entre le document et le regard porté sur lui qui peut produire, en situation, une représentation nouvelle* » (Jeanneret, 2000 : p. 119).** Cette posture a permis de revenir sur des objets qui sont au cœur des problématiques des sciences de l'information comme la notion de thème, le rôle de la terminologie ou encore la question de l'évaluation. Ainsi, dans le chapitre suivant, nous montrerons que l'indexation par thèmes doit non seulement prendre en compte les genres textuels (cf. la première partie de ce document) mais d'autres notions comme celle d'*événement* pour indexer des thèmes dans la presse. Nous mettrons aussi en évidence la pertinence du métadiscours pour consulter des documents scientifiques en SHS, relativisant ainsi le rôle de la terminologie pour l'indexation de l'information spécialisée. Nous poserons enfin la question de la contribution de l'émotion à l'évaluation de l'information médicale dans les forums de santé.
- Un troisième ensemble de méthodes réside dans la confrontation entre la caractérisation du contenu de certains documents (des thèses, des

journaux anciens, des manuscrits autographes), l'analyse des discours des usagers sur leurs pratiques informationnelles et sur l'usage de ces documents. Ces études contribuent à élaborer des préconisations pour guider l'indexation. Ces préconisations s'appuient sur les résultats d'enquêtes d'usage portant sur des dispositifs comparables, par exemple les usages des bibliothèques numériques (le projet CaNu XIX) ou sur des études d'usages et de pratiques informationnelles conduites sur une catégorie ciblée d'individus (les projets Métilde et SCIENTEX).

En définitive, l'intérêt de nos méthodes est qu'elles mettent à jour des formes de signification et des plans d'organisation de l'information qui tiennent compte des documents eux-mêmes ainsi que des discours des usagers sur leurs pratiques : elles participent d'une approche contextualisée de la recherche d'information et offre une contribution à l'organisation des connaissances.

3. Les discours pour guider l'organisation des connaissances : *lire, annoter, communiquer…*

Dans ce chapitre, nous envisageons les apports de notre recherche à l'organisation des connaissances. Nous nous appuyons sur l'analyse des discours, comme indiqué précédemment. Les discours révèlent des éléments d'information qui contribuent à organiser les connaissances. Ces éléments renseignent sur ce que recherchent les usagers, comment ils utilisent les dispositifs, quels sont les passages lus, les notes qu'ils prennent, comment ils interprètent et communiquent ce qu'ils sélectionnent.
Ces éléments sont le produit d'enjeux socio-culturels, disciplinaires ou professionnels qu'une approche anthropologique des connaissances saisirait dans leur globalité (cf. La revue *Anthropologie des Connaissances*). Ils contribuent à appréhender la recherche d'information comme processus social contextualisé.

3.1. L'évaluation n'est pas le seul enjeu de l'organisation des connaissances

Lorsque nous avons participé à des recherches finalisées, les demandes de nos commanditaires ont toujours été nettement orientées vers l'évaluation de leurs

dispositifs, avec un objectif qui était moins celui de remettre en question les choix techniques déjà réalisés que d'amener un, voire, des public(s) à les utiliser. Dès la première étude conduite au GRESEC – l'enquête auprès des médecins spécialistes en CHU (NOESIS) –, nous comprenions que les enjeux techniques et sociaux étaient étroitement imbriqués. Assez rapidement, les membres de l'axe impliqués dans le projet avaient choisi de se concentrer sur les pratiques des médecins hospitaliers face à l'affluence des ressources numériques, plutôt que sur la plate-forme de recherche d'information médicale *stricto sensu* :

> *« A l'origine, l'enquête présentée ici a été initiée en relation avec le projet européen NOESIS visant à développer un système complexe de bibliographie électronique médicale. L'objectif de notre étude a été de cerner les pratiques informationnelles des médecins du CHU dans leur globalité, de comprendre leurs comportements informationnels en relation avec les besoins ressentis dans l'activité professionnelle, d'identifier les facteurs de satisfaction et les freins rencontrés, et de faire ressortir les attentes éventuelles de cette population par rapport aux dispositifs informationnels électroniques. » (Staii et al., 2008 : p. 70)*

Ce choix délibéré nous conduisait collectivement à nous démarquer de la voie de l'évaluation des systèmes, l'approche purement technique des dispositifs étant trop restricitive. En outre, ainsi que le décrit très justement Joëlle Le Marec au sujet des études d'usage dans le champ culturel, si les concepteurs prennent le « *risque de voir éclater les catégories de l'offre au moment de l'analyse des usages* » (Le Marec, 2004 : p. 356), il n'est pas étonnant qu'ils renoncent à la possibilité d'utiliser les résultats à des fins d'aide de conception. Dans le cas de ce projet européen destiné à développer un premier prototype à Grenoble, il était prévisible que toutes les préconisations, même issues d'observations menées *in situ*, seraient lettre morte. Cette situation étant monnaie courante, Joelle Le Marec prône de manière alternative une réflexion sur la conception en termes de médiations (*ibid.,* p. 357).

Les recherches finalisées auxquelles nous avons participé n'écartent pas pour autant la technique, cette dernière n'étant jamais neutre. Sur ce point, plusieurs chercheurs, tel Adrian Staii, pensent les techniques comme des *« objets imprégnés de social »* (Staii, 2012 : p. 41). Ces auteurs mettent l'accent notamment « *sur le rôle des idéologies et de l'imaginaire (Flichy, 2001 ; 2004) dans la construction des techniques* » (Staii, 2012 : p. 42) et considèrent que les techniques « *sont également porteuses de*

contraintes (autant fonctionnelles stricto sensu que sociales au sens large) avec lesquelles les usagers sont obligés de composer […] » (*ibid*). En ce sens, l'analyse des dispositifs est éclairante pour comprendre la manière dont les connaissances sont organisées et comment les dispositifs s'intègrent aux pratiques existantes. Et c'est à ce titre que des préconisations en matière de conception peuvent être émises.

3.2. Il existe deux visions possibles de l'organisation des connaissances

L'organisation des connaissances peut être définie dans une perspective instrumentale ou processuelle. C'est ainsi que nous interprétons la définition que donne Birger Hjørland (2008) de l'organisation des connaissances :

> *« Knowledge Organization (KO) is about activities such as document description, indexing and classification performed in libraries, databases, archives etc. These activities are done by librarians, archivists, subject specialists as well as by computer algorithms. KO as a field of study is concerned with the nature and quality of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations and concepts. »*[152] *(Hjørland, 2008 : p. 86)*

Une vision instrumentale des connaissances est orientée vers la création de « *schémas d'organisation* » variés[153]. Ces schémas d'organisation, bien qu'ils soient conçus pour être intégrés à des systèmes d'information – qui ne sont pas uniquement informatiques –, sont alors analysés comme produits d'une culture, d'une époque, d'une communauté ou d'une société. De nombreuses recherches analysent les connaissances produites dans les organisations ou les institutions culturelles. Au sein de la Library of Information Science (LIS), les publications portant sur les classifications produites par et pour les bibliothèques, occupent une place très importante, ainsi que la question des normes et des standards. Les classifications sont

[152] « *L'organisation des connaissances concerne les activités telles que la description de document, l'indexation et la classification effectuée dans les bibliothèques, les bases de données, les archives, etc. Ces activités sont réalisées par les bibliothécaires, les archivistes, les documentalistes généralistes ou spécialisés ainsi que par les algorithmes informatiques. L'organisation des connaissances, en tant que champ d'étude, porte sur la nature et la qualité des processus d'organisation des connaissances aussi bien que sur les systèmes d'organisation des connaissances mis en œuvre pour organiser les documents, les représentations des documents et les concepts.* » (Hjørland, 2008 : p. 86) – Notre traduction

[153] Voir la citation complète en introduction de ce mémoire (Polity et al., 2005 : p.13)

de nature encyclopédique ou sectorielle et privilégient le partage de savoirs de type disciplinaire. En SIC, les publications abordant l'organisation des connaissances, s'intéressent à des connaissances issues du champ de l'information scientifique et technique (la santé, l'agriculture, la chimie, l'énergie, l'aérospatiale, etc.) et plus récemment d'autres secteurs économiques (la banque, le tourisme ou les services) cette évolution étant à mettre sur le compte de l'omniprésence de l'information professionnelle dans les organisations (Clavier et Paganelli, 2013).

Les enjeux scientifiques de l'organisation des connaissances sont souvent imbriqués avec ceux de la gestion des connaissances (Michel, 2001), de la veille terminologique (Ibekwe-San-Juan, 2006) ou de l'intelligence économique (Amos, 2005 ; Couzinet 2005) et induisent des effets de « brouillage » (Hudon et Mustapha el Hadi, 2013)[154]. La diversité des accès à l'information sur le web interroge les formes d'hybridation de langages d'indexation issus de classifications « formelles » *versus* « naïves » (Ihadjadene et Favier, 2008)[155], ainsi que la coexistence de plusieurs modèles d'organisation, hiérarchiques *versus* distribués (Hudon et Musatapha el Hadi, 2010). Elle interroge la structure des langages sociaux comme les folksonomies (Broudoux 2013 ; Durieux, 2010 ; Francis et Quesnel, 2007). D'autres travaux proposent des outils d'évaluation des systèmes d'organisation des connaissances (Zacklad, 2007) et abordent l'épineuse question de l'interopérabilité des systèmes (Zacklad, 2010).

Une vision processuelle de l'organisation des connaissances décrit les activités humaines ou automatiques en lien avec la production de connaissances à partir de différents supports (papier, numérique, chimique, optique, etc.), relevant de différentes sémiotiques (image fixe ou animée, texte, multimedia) et dans une infinité de domaines de connaissances ou d'activités sociales. Birger Hjørland souligne qu'il y a une vision étroite de l'organisation des connaissances et une vision large. La LIS

[154] « *[…] les relations existant entre champs d'activités et de recherche voisins mais à visées différentes, l'organisation des connaissances et la gestion des connaissances par exemple, semblent de plus en plus embrouillées.* » (Hudon et Mustapha el Hadi, 2013 : p. 10)

[155] « *Une classification formelle, effectuée à partir d'une structure imposée est qualifiée par Beghtol de "professionnelle" alors qu'une classification instinctive faite par les créateurs de documents eux-mêmes est considérée comme « naïve ». La « classification » doit être entendue au sens large d'activité de répartition en catégories d'une ensemble logiques ou réels* » (Ihadjadene et Favier, 2008 : p. 13)[155],

s'inscrit dans la première vision et se préoccupe des activités liées à la description de documents, à l'indexation et la classification produites dans les bibliothèques, les bases de données bibliographiques par des spécialistes, des ordinateurs ou des profanes (Hjørland, 2008)[156]. La seconde est une vision large de l'organisation des connaissances : la connaissance est « *au coeur de toutes les activités humaines, toutes celles du moins dans lesquelles l'esprit est de quelque manière engagé* » indique Jean Meyriat[157]. Pour Birger Hjørland, l'organisation des connaissances concerne la division sociale et mentale du travail qui relève notamment de la sociologie des sciences. Serge Agostinelli souligne que lorsqu'elles ne sont pas disciplinaires, les connaissances « *s'expriment dans le cadre d'une théorie holiste de l'homme* » (Agostinelli, 2013 : p. 66).

Viviane Couzinet relève des éléments qui illustrent les spécificités de notre discipline pour aborder l'organisation des connaissances. Elle pose ainsi la question de « l'exception française » des SIC (Couzinet 2006 ; 2012 ; Ibekwe-SanJuan, 2012). L'auteur considère qu'en « *véhiculant les référents collectifs et en donnant un ordre aux connaissances, les langages donnent à voir et transmettent une certaine vision du monde* » (Couzinet, 2012 : p. 46). Elle cite des exemples d'études où l'indexation est une pratique qui contribue à « la propagation des idées » dans le domaine de l'art (cf. Régimbeau). L'indexation est envisagée alors comme processus communicationnel (Courbières, 2002) ou « l'indicialité », sorte de « culture de l'indexation » est un « processus social et communicationnel de la désignation » entrant « pleinement dans le domaine de la culture » (Kovacs et Timimi, 2006). Quant aux langages, que ce soient des mots, des termes ou des concepts, ne sont-ils pas, ainsi que le suggère Muriel Amar toujours extraits de « corpus de textes », des « mots de discours »,

[156] *« In the narrow meaning Knowledge Organization (KO) is about activities such as document description, indexing and classification performed in libraries, bibliographical databases, archives and other kinds of "memory intuitions" by librarians, archivists, information specialists, subject specialists, as well as by computer algorithms and laymen. KO as a field of study is concerned with the nature and quality of such knowledge organizing processes (KOP) as well as the knowledge organizing systems (KOS) used to organize documents, document representations, works and concepts. Library and Information Science (LIS) is the central discipline of KO in this narrow sense (although seriously challenged by, among other fields, computer science).* »(Hjørland, 2008)

[157] cité par Viviane Couzinet (2006)

(Amar, 2000 : p. 174) l'indexation consistant moins à partager des mots, qu'à leur assigner une place dans l'univers des discours (*ibid*., p.176) ?

3.3. Différents éléments influencent l'organisation des connaissances

Dans la première partie de ce mémoire, nous avons considéré l'organisation des connaissances comme des processus automatiques à l'œuvre dans des artefacts techniques. La dimension instrumentale était bien présente, les dictionnaires électroniques étant des « systèmes d'organisation des connaissances ». En outre, les annotations morphosyntaxiques encodées en XML participaient de la « représentation de l'organisation des connaissances » (Gnoli, 2012 : p. 52)[158]. La réflexion sur l'organisation des connaissances était donc guidée par des enjeux de conception. La seconde partie en revanche, ne poursuit pas un tel objectif. Nous ne produisons pas de systèmes d'organisation des connaissances et nous ne développons pas de méthodes de traitements automatiques : **nous recueillons dans les discours des éléments d'information qui contribuent à organiser des connaissances. Le protocole de collecte des corpus garantit que les discours analysés s'inscrivent dans une « vision étroite »[159] de l'organisation des connaissances.** Par exemple, lorsque les médecins spécialistes en CHU disent rechercher des schémas pour faire leur cours, cette indication est importante pour cerner la nature de connaissances à indexer (quels types de schémas sont privilégiés par exemple). Ainsi, en prenant en compte la manière dont les usagers s'informent, quels outils ils utilisent, ce qu'ils font de l'information, à quoi elle leur sert, et pourquoi ces façons de faire sont révélatrices de pratiques ancrées socialement, nous nous dotons de méthodes pour identifier les connaissances utiles aux usagers, les caractériser et les organiser. Elles se révèlent influencées par divers éléments qui varient suivant les contextes de recherche d'information : l'environnement de travail, le type de tâches à réaliser, les contraintes techniques, linguistiques, les logiques de conception des Tic, les contenus.

[158] Claudio Gnoli considère l'organisation des connaissances comme un discipline comportant quatre niveaux : la théorie, les systèmes, la représentation, les applications. (*ibid.* p. 52)

[159] Au sens de Birger Hjørland (cf. *supra*)

Pour illustrer notre propos, nous revenons sur des études de cas présentées précédemment en mettant l'accent sur les résultats des enquêtes d'usages et de pratiques. La lecture que nous faisons de ces enquêtes est guidée par la recherche d'éléments saillants qui jouent un rôle pour l'organisation des connaissances. Nous nous intéressons dans la suite aux projets NOESIS et Métilde.

3.3.1. NOESIS : la terminologie médicale, les schémas et les articles de revue

L'enquête sur les pratiques informationnelles des médecins spécialistes en CHU révèle trois types d'activités liées à l'information spécialisée : la recherche, l'enseignement et la pratique médicale, cette dernière étant plus minoritaire (Staii et al., 2008 : p. 82). Les médecins ont par ailleurs l'obligation de se tenir informé régulièrement. Ainsi, la formation continue relève-t-elle d'une disposition légale présente dans le code de déontologie médicale et officiellement sanctionnée par des textes législatifs et réglementaires : la documentation, la participation à des séminaires, des conférences, etc. font partie de ce dispositif. En outre, la formation documentaire constitue un passage obligé du cursus universitaire. Par exemple à Grenoble, dès la seconde année, les sources documentaires incontournables sont présentées : la base Pubmed, le MeSH[160], les catalogues des bibliothèques de médecine, les banques de données en santé publique, etc. en lien avec des travaux dirigés en recherche d'information. Ultérieurement, dans le cadre de leur spécialisation, les médecins sont confrontés à une très grande diversité de ressources, parmi lesquelles figurent les revues spécialisées. Cette population se caractérise par une activité informationnelle très soutenue. Elle fait essentiellement usage de TIC, privilégie les circuits d'information validée. Les revues spécialisées en anglais sont les sources les plus valorisées et les médecins spécialistes sont très bien informés, ainsi que nous l'indiquions dans notre article commun :

> *« […], il apparaît que les enseignants chercheurs en médecine lisent en règle générale plus que la moyenne des scientifiques [TENOPIR et al. 2003b], caractéristique qui se retrouve aussi sur le terrain de la pratique professionnelle (traditionnellement, les médecins pratiquants lisent plus que les physiciens ou les ingénieurs par exemple - [WILLIAMSON et al., 1989]). Cette population*

[160]Medical Subject Headings (MeSH) développé par la National Library of Medicine : http://www.nlm.nih.gov/mesh/MBrowser.html

> *préfère également de plus en plus les sources électroniques [DeGROOTE et DORSCH, 2001], et elle semble être également attachée aux abonnements personnels [TENOPIR et al. 2003b] à des revues de référence. Parmi les sources d'information les plus valorisées, les revues spécialisées occupent d'ailleurs de loin la première place, les médecins étant très attachés à des circuits et à des sources validées qui évitent les pertes de temps et les vérifications croisées [TENOPIR et al. 2003b]. » (Staii et al., 2008 : p. 78)*

Par conséquent, l'organisation des connaissances est influencée par ce contexte pédagogique et professionnel ainsi que par l'offre disponible en matière de ressources. Le marché économique des bases de données médicales est, depuis la seconde guerre mondiale, dominé par les Etats-Unis – la première base de données bibliographique MEDLARS[161] remonte aux années 40, suivie de MEDLINE (MEDLARS on line) et de PubMed. Les systèmes d'organisation des connaissances à l'oeuvre dans PubMed héritent les principes d'organisation du MeSH, le thesaurus utilisé dans ces bases de données. Les descripteurs renvoient à une terminologie hyperspécialisée exprimée en langue anglaise[162], organisée de manière hiérarchique et la consultation de ces bases de données nécessite une connaissance approfondie non seulement du domaine, mais aussi de ce langage documentaire. L'accès aux sources est donc conditionné par une acculturation forte à la littérature scientifique diffusée par ces grandes bases de données, aux normes de communication anglo-américaines qui structurent ce champ international de la communication scientifique et aux langages utilisés pour représenter l'information.

Les revues spécialisées constituent le type de document privilégié par les médecins.

> *« Les sujets sont tous capables de citer les revues importantes de leur spécialité et ils les lisent tous régulièrement. La quasi-totalité de ces revues sont en anglais. En plus des revues relevant de leur spécialité, la majorité des sujets (82%) lit d'autres revues scientifiques, médicales ou généralistes (New England, Nature, Science). » (ibid. p. 82)*

Généralement sélectionnés après consultation des revues de sommaires, les articles scientifiques sont le type de document le plus consulté. Ils font l'objet de plusieurs stratégies de lecture qui mettent toutes en évidence le rôle fondamental de la

[161] Medical Literature Analysis and Retrieval System.

[162] Bien qu'il existe une version française du MeSH http://mesh.inserm.fr/mesh/new_desc.htm

structure hiérarchique du texte pour accéder au contenu. En sciences expérimentales, la structure hiérarchique IMMRaD coïncide avec la structure rhétorique, si bien que les lecteurs ont des parcours assez stéréotypés : lecture du résumé, de l'introduction de la conclusion, puis éventuellement des matériel et méthode, ou de la discussion. Les documents jugés intéressants sont imprimés et éventuellement annotés, puis ils sont classés. Dans le cadre de l'activité d'enseignement, les schémas sont des informations très exploitées. L'annotation de la bibliographie reste individuelle ou cantonnée au cercle limité d'une équipe de recherche, les médecins spécialistes, également chercheurs, organisant des séances bibliographiques avec les doctorants.

3.3.2. Métilde : les variantes formelles, le support papier et le livre

Les pratiques informationnelles des spécialistes de littérature apparaissent aux antipodes des pratiques des médecins sur la question des langages, des types de documents consultés, des dispositifs informationnels utilisés, des normes de publication scientifique et des objectifs de l'annotation. Bien qu'elles se déroulent dans un cadre professionnel, les activités informationnelles des spécialistes des manuscrits d'auteurs du 19ème siècle relèvent d'une culture très différente[163].

L'activité de recherche est par tradition individuelle et les pratiques de publication valorisant le livre au détriment des articles scientifiques. En outre, le livre est aussi, et surtout, un objet d'étude : le livre, c'est l'Oeuvre. Qu'il s'agisse d'étudier une œuvre dans les livres édités, ou la genèse d'une œuvre dans les manuscrits d'auteurs, les études littéraires appréhendent toujours les documents comme objet d'étude et objet de recherche. Le fait que ces objets se confondent est très important pour l'organisation des connaissances – et également pour le développement d'un dispositif technique – puisque plusieurs enjeux se superposent rendant caduque la distinction classique entre tâche principale et tâche secondaire, c'est-à-dire, où la recherche d'information est secondaire par rapport à l'activité principale. Les manuscrits sont en effet le matériau central de l'édition critique et de la génétique

[163] Nous rendons compte de conclusions qui n'ont pas encore été publiées par les chercheurs de ce projet et qui ont cependant été présentées lors de séminaires à nos partenaires. Des publications sont actuellement en cours ainsi qu'une Journée d'études programmée en avril 2014.

textuelle. Les spécialistes font un travail de transcription qui donne lieu à des éditions critiques. Ces transcriptions sont de différente nature : transcription linéarisée, diplomatique, pseudo-diplomatique, etc. Du point de vue éditorial, les transcriptions sont des annotations critiques. Jusqu'à présent, le circuit traditionnel pour l'édition critique était, et est toujours, le papier, mais depuis quelques années, de nombreux projets d'édition numérique se développent. Il s'ensuit de nouveaux objectifs relatifs à l'édition en ligne, comme la valorisation du patrimoine littéraire, la communication scientifique, le développement d'outils de techniques et de méthodes pour accompagner les projets d'humanités numériques.

Si l'on se limite à l'observation des pratiques informationnelles[164], il apparaît que les spécialistes de littérature ont une pratique essentiellement orientée vers l'étude de documents papier. Certes, les spécialistes en littérature utilisent des dispositifs techniques : par exemple les catalogues collectifs, tel le catalogue général des manuscrits (CGM) ou Calames, etc. Cependant, à l'inverse des médecins qui désertent les bibliothèques, les chercheurs en littérature fréquentent assidument les fonds spécialisés des bibliothèques ou les archives. Les spécialistes des manuscrits d'auteurs se comptent sur le doigt de la main. Ainsi, parle-t-on d'une dizaine de spécialistes de Stendhal dans le monde, notamment au Japon et aux Etats-Unis.[165] Pour ces spécialistes, se pose la question de l'accès à ces documents précieusement conservés dans les bibliothèques françaises et italiennes. C'est ainsi que la mise en ligne des manuscrits d'auteur est considérée comme une grande opportunité.

Que recherchent alors les spécialistes de littérature dans les manuscrits d'auteurs ? Pour les chercheurs en littérature Françoise Leriche et Cécile Meynard, les *Journaux et Papiers* de Stendhal appartiennent à la catégorie des *« manuscrits de travail »* qui sont des *« documents autographes où le texte s'élabore peu à peu, de ratures en reprises »* (Leriche et Meynard, 2008, note 8, p.11). Ces auteurs indiquent que les manuscrits de travail ont longtemps été négligés par les universitaires qu'ils considéraient comme des brouillons « illisibles », et lorsqu'ils y avaient recours,

[164] Enquête en cours réalisée par Evelyne Mounier et Céline Paganelli.

[165] Ces informations nous ont été communiquées oralement par Cécile Meynard lors de séminaires au laboratoire Traverses 19-21.

c'était pour s'intéresser aux « dernières étapes », aux « épreuves corrigées », c'est-à-dire les manuscrits « au net », afin de fournir les « variantes » des éditions savantes. Les auteurs mentionnent qu'il a fallu attendre le développement des études de génétique dans les années 60-70, pour que les « étudiants avancés » et les « spécialistes » analysent les archives de la création littéraire dans les bibliothèques. L'édition génétique résulterait, selon ces auteurs, « d'une nouvelle ambition éditoriale : éditer non pas des textes mais des « brouillons en expansion » et élargir l'accès à ces documents jusque là réservés à un public confidentiel.

La question qui se pose pour les spécialistes de littérature serait donc bien de montrer le « processus d'écriture » (Grésillon, 2006), de montrer « l'inachevé » de montrer le « texte en mouvement » (Leriche et Meynard, 2008 : p. 12). Mais que veut-on précisément *montrer* ? Est-ce seulement « rendre visible » des documents sur le web ? Est-ce identifier les éléments propres aux manuscrits ? Est-ce renseigner sur la « dynamique d'écriture » de l'auteur (Lebrave, 2007)? Est-ce apprendre à lire, déchiffrer et interpréter, pointer sur ce qui fait sens dans un manuscrit et mettre en évidence le « long travail de déchiffrement » du généticien cherchant à « domestiquer la trace graphique » ? (Lebrave, 2007).

3.3.3. Bilan : des langages hiérarchisés aux marques formelles spatialisées

Les systèmes d'organisation des connaissances sont une composante essentielle des dispositifs informationnels. Joseph Tennis, professeur assistant à *l'Information School* de l'Université de Washington[166], indique que l'étude de ces systèmes constitue une grande partie de la littérature relative à l'organisation des connaissances (Tennis, 2013 : p. 17), et que le « poids du langage » y est central. **Ainsi, nous pouvons identifier un premier élément d'influence qui est le rôle des langages dans les systèmes d'organisation des connaissances médicales.**

Dans le cas de l'information médicale, l'ordre des connaissances mis en œuvre dans les systèmes de classifications hiérarchiques comme le MeSH est calqué sur l'organisation des savoirs scientifiques par discipline. La terminologie apparaît

[166] http://ischool.uw.edu/people/faculty

comme le point nodal des systèmes d'organisation des connaissances et elle impose à tous les acteurs, professionnels de l'indexation et usagers, une veille permanente. Le choix des termes, leur recueil, leur représentation soulève essentiellement des questionnements relatifs à la problématique du signe et de son expression sous la forme de concepts, c'est-à-dire au caractère non bijectif entre les langues naturelles et les langages documentaires. Toutefois, l'étude de cas sur les manuscrits de Stendhal indique qu'un système d'organisation des connaissances ne peut se limiter aux langages issus du mode verbal[167].

Nous pouvons identifier un second élément d'influence qui est le rôle des marqueurs sémiotiques de « la rature » dans les systèmes d'organisation des connaissances des manuscrits autographes. Nous entendons par là toutes les variantes formelles qui ponctuent les brouillons, qui sont codifiées dans les transcriptions, mais pas forcément représentées dans les index[168] Ainsi, les spécialistes de l'édition critique et de la génétique textuelle s'intéressent-ils à tous les éléments qui participent de la description de la page, laquelle constitue l'unité de travail du chercheur. Citons sans prétention d'exhaustivité, le rendu formel de la page, ce que les spécialistes nomment le « découpage topographique du texte » (Lebrave, 2007), les ratures, les biffures, les becquets d'insertion, les surcharges, les ajouts, les suppressions, la ponctuation, mais aussi l'aspect du papier, son grain, son format, sa couleur, les traces de colle ou de reliure, les taches diverses, la couleur de l'encre, qui indique le temps de séchage, le tracé qui renseigne sur le scripteur et le rythme d'écriture, etc. Cet objectif met au premier plan la dimension graphique et formelle des textes, plutôt que le langage, elle met également en évidence les dimensions propres au support physique. Contrairement aux langages classificatoires, ces connaissances ne sont pas hiérarchisées, elles sont spatialisées : dans les *Journaux et Papiers* de Stendhal, la page est l'unité de transcription, elle doit être l'unité de référence pour organiser les connaissances.

Ensuite, l'organisation des connaissances concerne également la nature des documents consignés dans les dispositifs. Dans le cas des médecins, les habitudes

[167] En attente de résultats définitifs sur les enquêtes de pratiques.

[168] Dans la plate-forme Clélia, l'on peut rechercher les biffures.

de publications scientifiques en anglais et le rôle central que joue l'article de revue scientifique, sont déterminants. Pour les spécialistes de littérature en revanche, la consultation de documents sur support papier dans les bibliothèques était, et est souvent encore, la pratique. Ainsi, il est raisonnable d'imaginer que tous les dispositifs numériques qui favorisent la lecture « à la loupe » et qui accompagnent le travail d'annotation (individuel) des spécialistes seront adoptés.

Enfin, les enquêtes sur les usages et les pratiques informationnelles, bien que n'ayant pas pour finalité immédiate d'évaluer les dispositifs, pointent certains éléments pouvant favoriser ou freiner l'adoption de ces outils, ainsi que nous l'indiquions dans l'avant-propos du rapport de recherche sur NOESIS :

> *« Cette étude ne cherche [donc] pas à cerner les attentes ou les pratiques des médecins en relation directe avec les outils Noésis, mais elle permet d'identifier des facteurs qui pourraient favoriser ou freiner l'adoption de ces outils et faciliter ainsi d'éventuels ajustements. » (Balicco et al., 2007 : p. 3)*

Ainsi, la plate-forme NOESIS qui reposait sur une logique d'accès au MeSH, avait toutes les chances d'être adoptée. En revanche, le module d'annotation imaginé par les concepteurs pour permettre aux experts d'annoter en ligne les articles de revues, avait de fortes chances de ne pas être utilisé. Il était prévu que ces annotations soient indexées et traduites dans toutes les langues. Or, le manque de temps des spécialistes, la culture du secret qui entoure cette profession, les modes de communication entre experts qui se déroulent plutôt sur le terrain des congrès, etc. tous ces éléments laissaient penser que ce module ne serait pas utilisé. D'une manière générale, les dispositifs informationnels ne modifient pas ou peu les pratiques existantes.

3.4. L'analyse des discours contribue à faire émerger des connaissances

L'organisation des connaissances ne s'est pas posée frontalement dans nos travaux en termes de champ d'étude. Cependant, les méthodes que nous utilisons pour décrire les contenus et sélectionner les informations en fonction des pratiques informationnelles des usagers, constituent une étape indispensable pour faire émerger des connaissances. De ce point de vue, les SIC ont un rôle à jouer qui se situe

à l'interface des techniques linguistiques d'extraction de connaissances, des méthodes d'entretiens issus des SHS et de l'ingénierie des connaissances.

Nous revenons sur deux études de cas qui concernent l'indexation, avec en arrière-plan la question de l'automatisation des traitements : les projets CaNu XIX et SCIENTEXT.

3.4.1. CaNu XIX : l'indexation de thèmes s'appuie sur les événements

Dans cette étude consacrée à la valorisation du patrimoine numérique de presse nous nous sommes intéressée au journal hebdomadaire lyonnais le *Progrès Illustré* (1890-1905) numérisé et mis en ligne par la bibliothèque municipale de Lyon sur un site expérimental, considéré comme pilote en France (Landron, 2010). Ce site, aujourd'hui disparu, a été remplacé par une bibliothèque numérique consacrée à la presse lyonnaise de 1790 à 1944, dans laquelle figurent plusieurs titres dont le *Progrès Illustré*[169]. En 2006, la bibliothèque municipale de Lyon engagée comme bien d'autres institutions culturelles dans une politique de numérisation de masse, souhaitait mettre en accès libre ces archives au grand public et privilégiait la réflexion sur la valorisation : la « construction de parcours thématiques par les professionnels des bibliothèques » était un objectif de valorisation imposé par le programme CaNu XIX.

3.4.1.1. Diversifier les modes d'accès aux collections de presse

Nous nous sommes emparée de la notion de parcours thématique destiné à faire découvrir les collections au grand public Nous appelions *parcours thématique* « un dispositif de mise en exposition d'une collection numérique suivant un ensemble de *sujets* prédéfinis. » (Clavier, 2010 : p. 102). Nous proposions que ce dispositif soit complémentaire de la recherche plein texte et des dossiers thématiques. Le rapport Teissier (2010) consacré à la numérisation du patrimoine écrit, préconisait en effet un accès plein texte afin de favoriser le référencement des bibliothèques numériques par les moteurs de recherche. Mais, pour pouvoir poser une requête, encore faut-il

[169] http://collections.bm-lyon.fr/presseXIX/showObject?id=PER003&date=00000923

que les usagers sachent ce qu'ils cherchent. Or, l'étude d'usages et de pratiques conduite par les collègues du projet CaNu XIX était formelle sur ce point : si certains publics ont des besoins bien précis, comme les historiens par exemple, une grande partie des usagers comme le « grand public lettré », ne cherche rien de particulier :

> *« Enfin, nous rencontrons une forte proportion d'usagers qui n'ont pas de recherche particulière à effectuer et qui arrivent là par hasard. Il ne s'agit donc pas d'un public habituel usager de la presse ancienne, mais c'est la mise en ligne qui commence à attirer un public différent. » (Paganelli et al., 2011b : p. 255)*

Pour ce motif, plusieurs professionnels de l'information affichent une certaine prudence à l'égard de l'accès plein texte et préconisent d'autres formes d'accès au contenu :

> *« On touche là à la nécessaire réflexion sur l'accès aux documents, qui ne saurait calquer les catalogues des bibliothèques : il faut donner à voir une collection numérique avant d'offrir une recherche précise, les promenades libres ou guidées ont une signification. » (Westeel, 2009 : 30)*

Inversement, le mode d'accès par les dossiers thématiques favorise pleinement la découverte préalable des collections et les bibliothèques numériques en font d'ailleurs grand usage :

> *« La conception de dossiers thématiques est une pratique largement répandue sur les sites web des bibliothèques numériques (Gallica) ou des agrégateurs de contenus numériques (Europeana) : ils permettent de faire vivre le site, de l'animer et également de mettre en valeur un patrimoine numérique suivant un choix de thèmes attractifs. » (Clavier, 2010 : p. 102)*

La bibliothèque numérique de Lyon propose d'ailleurs sur sa page d'accueil plusieurs dossiers thématiques consacrés à la bicyclette, l'anarchisme, les grandes affaires criminelles, la mode etc. Les dossiers reposent sur le principe de la synthèse et donnent lieu à un nouveau document placé sous la responsabilité d'un auteur. Ils apparaissent dans un espace dédié, actuellement une rubrique. Dans la précédente version du site, ils figuraient dans une fenêtre avec une boîte à listes distincte du kiosque et de l'interface de recherche, comparable à un espace d'exposition, au sens muséal du terme. Bien que fort attractifs, ces dossiers présentent néanmoins une limite de taille : ils supposent un travail humain important qui nécessite une très

bonne connaissance des collections et du contexte historique. Bref, ces dossiers sont le fruit de spécialistes.

Aussi, la proposition que nous avions émise visait à créer des parcours thématiques reposant sur le double principe d'une indexation automatique du texte intégral afin de limiter le travail humain, et sur la caractérisation de contenus définie en fonction des centres d'intérêt des publics identifiés.

3.4.1.2. Des enquêtes d'usages pour cerner les publics et les attentes

En nous appuyant sur les enquêtes consacrées aux pratiques de recherche d'information sur internet (Matharan et al., 2008), aux usages de bibliothèques numériques comme Gallica[170] ou Europeana[171] ainsi que sur les résultats de l'enquête conduite par les membres du projet CaNu sur les usages du *Progrès Illustré*, nous relevions des éléments guidant le choix des connaissances à représenter pour répondre aux attentes de ces publics.

En premier lieu, l'accès aux ressources internet semblait facilité lorsque l'information était organisée sous la forme de « thèmes précis » (Marathan et al. 2008) :

> *« Une question ouverte permettait aux répondants d'indiquer quels sites ils fréquentaient (sur Internet) et auxquels ils participaient le plus. 240 réponses ont en tout été récoltées. Une logique thématique nette en émerge : on voit que les répondants mentionnent avant tout des sites traitant d'un thème précis, spécialistes d'un sujet (Matharan et al., 2008 : 7) » (Clavier, 2010 : p. 104)*

Des études confirmaient que la navigation par thèmes dans une collection de documents – donc dans un environnement fermé – évitait la désorientation :

> *« Ce mode d'accès, qui permet de naviguer dans une collection suivant une logique thématique ou par sujet, est considéré comme beaucoup plus performant que le mode de recherche par mot-clé (Abdullah et Gibb, 2009 cité par Da Sylva, 2009). Pierre Zweigenbaum et Benoit Habert (2004) indiquent que le « foisonnement de données textuelles et d'outils » conduit la plupart du*

[170] (Lesquins, 2007).

[171] (Bouvier-Ajam, 2007).

> *temps à une désorientation des usagers, et qu'il est nécessaire de fournir des « boussoles sémantiques » pour naviguer dans les documents. » (ibid.)*

… et que ce mode de navigation permettait une « appropriation graduelle du contenu » favorisant après coup, une prise en main de l'interface de recherche en mode texte :

> *« Plusieurs travaux montrent que la navigation dans une structure pré-établie serait une aide considérable pour les usagers. Il existe plusieurs dénominations pour qualifier ces outils : « outil de butinage » pour Da Sylva (2009), il permettrait de guider l'usager et favoriserait « une appropriation graduelle d'un contenu, même si l'utilisateur n'a aucune connaissance préalable de celui-ci. » ; il serait également une « aide à la lecture » qui permettrait l'évaluation de la pertinence de documents. « Système de visualisation de l'information » pour Davis (2006)[172], il permettrait de regrouper les documents semblables, de cerner la pertinence des documents retrouvés et de combiner la recherche par mots-clés. » (ibid.)*

En second lieu, il ressortait de l'enquête conduite par Evelyne Mounier, Céline Paganelli et Stéphanie Pouchot (Paganelli et al, 2011 a et b) que le lecteur-type de ces archives de presse souhaitait effectuer des recherches de dates, de lieux, de personnes et d'événements ainsi que nous le rapportions dans notre article :

> *« Le profil type du lecteur de la version en ligne du Progrès Illustré est lettré, souvent à la retraite, peu préoccupé par la pertinence des résultats fournis, parce qu'il ne cherche rien de précis. Ce qui le gêne en revanche, ce sont les problèmes d'affichage, les caractères étant souvent trop petits. Il s'intéresse essentiellement aux dates, aux lieux, aux événements, aux personnes, en dernier lieu aux sujets. Cette enquête ne révèle pas d'engouement particulier pour les parcours ou les dossiers thématiques. » (Clavier, 2010 : p. 105)*

Au vu de ces résultats, nous préconisions deux objectifs : favoriser l'accès aux collections par thèmes et permettre la recherche de lieux, de dates, de personnes et d'événements.

[172] cité par Da Sylva (2009 : 264)

3.4.1.3. Définir les thèmes et les événements en tenant compte de la morphologie du journal

En posant la question des thèmes et des événements dans la presse, nous ouvrions un vaste chantier de recherche. Dans notre article, nous posions les termes de cette recherche qui mériterait de plus longs développements.

La notion de thème est au cœur des préoccupations des disciplines littéraire, linguistique et des pratiques documentaires. L'une des raisons qui fait obstacle à une définition synthétique du *thème* réside dans l'extrême hétérogénéité des perspectives d'analyse, le thème apapraissant dans le cadre de la phrase, du texte et du discours. En outre, les unités thématiques peuvent être identifiées syntaxiquement grâce à des « marqueurs de thématisation » ou être assimilées à des unités lexicales structurées en champs sémasiologiques ou onomasiologiques. Elles se manifestent dans des phrases qui composent un paragraphe ou dans un texte ou encore, ne correspondre à aucun énoncé mais à un « topic » ou une relation « d'à-propos ». Enfin, François Rastier (1996) indique qu'un thème peut renvoyer « *à une structure stable de traits sémantiques, récurrente dans un corpus, et susceptible de lexicalisations diverses.* » Il précise en outre que « *selon les discours et les genres, les normes de lexicalisation des thèmes varient* ». Concernant la perspective documentaire, nous indiquions que la notion de *thème* est mobilisée pour décrire des outils d'accès au contenu, les index thématiques, ou pour concevoir des documents de synthèses, les dossiers thématiques. Nous indiquions que la terminologie se révèle flottante entre *sujet* et *thème* : ainsi, les index thématiques permettant d'accéder aux documents *qui parlent de la même chose*, c'est-à-dire, qui traitent du même *sujet*.

Nous en venions à la question frontale de la constitution d'index thématiques dans un environnement numérique, ces derniers devant être intégrés aux documents primaires afin de servir d'outils de recherche et de lecture :

> *«L'auteur [Murielle Amar] indique que les index doivent permettre de manipuler non plus l'intégralité d'un document mais aussi des segments pouvant, le cas échéant, être combinés pour produire de nouveaux documents, sous réserve que soient introduites des connaissances contextuelles, externes au document. (Amar, 2004 : 62-63). » (ibid.)*

Or, Muriel Amar (2004) soulignait que cet objectif supposait une refonte profonde des objectifs de l'indexation documentaire, c'est-à-dire l'indexation en langage contrôlé. En effet, l'auteur considérait que la construction d'un thème – et corrélativement son indexation – était une opération fondamentalement discursive qui nécessitait des connaissances extérieures au document, auxquels les « utilisateurs » n'avait pas accès. Il en résultait, selon elle, une interprétation difficilement « reconstituable ». En outre, l'auteur mentionnait que les formulations utilisées dans les langages contrôlés sont de nature lexicale et référentielle alors que les thèmes ne sont pas des unités référentielles mais discursives. Ce constat « désespérant » conduisait l'auteur à poser la question de savoir comment concilier la thématisation et la référenciation dans les objectifs de l'indexation professionnelle. Partageant le même point de vue sur le caractère difficilement « discrétisable » des unités thématiques en catégories référentielles, nous retenions les propositions de Muriel Amar relative à l'indexation thématique. L'auteur prônait une indexation favorisant un accès direct au texte intégral, qui ne s'attache pas au lexique mais au discours. **Elle nommait « indexation discursive » le processus qui consiste à donner à l'utilisateur non pas « *des mots pour dire les thèmes* », mais « *des documents, regroupés thématiquement* », qui sont les contextes *« qui rendent intelligibles et interprétables les thèmes des documents » (ibid. 65).* Pour l'auteur, indexer consistait alors à permettre la construction des unités d'interprétation que propose le texte, et non les nommer.**

Cette conception de l'indexation était, selon nous, une voie prometteuse pour atteindre les événements et nous pensions que les « documents » regroupés thématiquement, donneraient accès aux événements. En effet, dans la presse, les thèmes et les événements sont des notions liées ainsi que le soutiennent des spécialistes de la linguistique textuelle :

> *« Ainsi, Jean-Michel Adam et Gilles Lugrin cherchant à typer les unités rédactionnelles et catégorielles de la presse contemporaine évoquent les thèmes, pour désigner « des objets de discours inséparables des familles d'événements » (Adam et Lugrin, 2000 : 13). Ils s'appuient sur les travaux de Maurice Mouillaud et Jean-François Têtu (1989) pour qualifier les « familles événementielles » de catégories référentielles apparaissant au sein des rubriques. Les nouvelles*

> *politiques, les catastrophes, les conflits sociaux sont, pour ces spécialistes des médias, des familles d'événements […] » (Clavier, 2010 : p. 108)*

Ainsi, contrairement aux pratiques documentaires qui indexent les thèmes sans tenir compte du genre textuel ou de la structure de l'information nous proposions au contraire, de tenir compte d'éléments qui participent de la morphologie du journal :

> *« Parmi les différentes dimensions qui interviennent pour le typage d'un thème, il y a notamment les éléments qui participent de la morphologie du journal, les rubriques qui permettent d'établir l'identité énonciative des journalistes, le péritexte (titres et intertitres), les genres, et, pour reprendre la proposition de Maurice Mouillaud et Jean-François Têtu, les « familles événementielles » (ibid.)*

Les éléments évoqués par Maurice Mouillaud et Jean-François Têtu (1989) ne sont pas des « documents » au sens d'entités documentaires autonomes, mais des éléments qui participent de la description d'un titre de journal. Nous considérions que les familles d'événements mentionnées parces auteurs faisaient partie de ces éléments. Nous indiquions que la qualification d'événement dans les médias n'était pas du ressort de la linguistique mais procédait d'une reconfiguration de la réalité « *déformée* » par *« l'industrialisation des métiers de la presse, le développement des technologies modernes de communication et/ou les intérêts économiques et financiers des groupes qui les fabriquent* » (Arquembourg, 2006 : 14). Nous indiquions en outre qu'il existait des typologies d'événements dressées dans le cadre de la norme de métadonnées IPTC[173] et qui permettaient de « ventiler l'actualité » (Palmer, 2006 : 53).

3.4.1.4. Bilan : une organisation des connaissances à trois niveaux

Cette prise en compte de la structure formelle et informative du journal inscrivait l'organisation des connaissances dans l'optique préconisée par Roger T. Pédauque qui définit le document à la lumière du numérique (Pédauque, 2006). Ainsi, dans l'étude

[173] L'IPTC (International Press and Telecommunications Council) est une organisation internationale créée en 1965 pour développer et promouvoir des standards d'échange de données à destination de la presse. http://www.iptc.org/cms/site/index.html?channel=CH0086<

CaNu XIX, nous proposions que les connaissances « utiles » à l'indexation de parcours thématiques devaient s'appuyer sur trois approches du document : le signe (le rôle des événements), la forme (la morphologie du journal) et le médium (la lecture de l'actualité). En outre, en raison du caractère imbriqué de la thématique avec les genres journalistiques, les types d'événements, les rubriques, la ligne éditoriale, etc. il nous semblait impossible de concevoir une indexation sur un texte à plat. Nous proposions une indexation sur trois niveaux.

Le premier niveau, destiné à choisir un sujet de parcours, et avait pour objectif de construire un index des *sujets*. Nous indiquions que ce choix était lié à la politique documentaire de l'institution en charge de la valorisation de la collection :

> *« Le choix d'un sujet ne repose pas sur une analyse des fréquences « des mots » des textes. Les critères de choix sont extérieurs au corpus : ils peuvent résulter de la connaissance qu'ont les indexeurs des usagers du fonds, de l'analyse des traces de requêtes, de spécificités jugées amusantes ou curieuses de la collection (les anecdotes, les caricatures), de préconisations de spécialistes sur l'apport de ces documents pour l'histoire, la littérature, ou la presse, de choix muséographiques (faire une exposition), etc. La liste des sujets n'est donc pas définie a priori, elle se construit au fil du temps, en fonction de choix de valorisation. Cette démarche est également celle qui préside à la construction des dossiers thématiques. » (Clavier, 2010 : p. 111)*

Le deuxième niveau, destiné à construire un index des *événements*, portait sur la définition même de cette notion. Nous indiquions qu'un événement comportait une dimension médiatique et historique : ce sont des sources extérieures qui permettent de leur attribuer un statut événementiel. Bien que l'indexation des événements soulève de nombreuses questions seulement esquissées dans notre article (comme par exemple, le fait que des événements ont pu être considérés comme tels au 19ème siècle, mais plus aujourd'hui), la notion nous semblait fondamentale pour caractériser la thématique, puisque c'est à travers les commentaires des événements que transparaissait le positionnement éditorial du journal. Nous indiquions que le lien entre les sujets et les événements pouvait se faire dans le cadre d'une approche actantielle : « Tel *sujet* peut figurer en position d'*actant* dans le cadre d'un procès qui décrit un *événement*. » (ibid., p. 112).

Le troisième niveau, destiné à l'interprétation, mettait en évidence les contextes qui « rendent intelligibles et interprétables les thèmes du document », suivant la proposition de Muriel Amar. Pour nous un thème est un positionnement éditorial sur l'actualité, c'est-à-dire les manifestations de « l'angle journalistique », au sens où l'angle peut être appréhendé en tant que « formant » (Ringoot et Pobert-Demontrond, 2004 : 108). Nous suivions en cela les propositions de Roselyne Ringoot et Philippe Robert-Demontrond sur l'analyse d'un thème informatif dans la presse :

> *« […] l'analyse d'un thème informatif nécessite un travail de diagnostic éditorial qui contextualise le traitement d'une information en fonction de la politique éditoriale d'un journal. » (Ringoot et Robert-Demontrond, 2004 : p. 88).*

Nous donnions un exemple de cet emboîtement entre la ligne éditoriale, les sujets et les événements dans l'analyse d'une rubrique : les Causeries.

> *« Par exemple, l'une des façons de révéler l'approche populaire du journal consiste à montrer que la voix du chroniqueur, dans l'analyse qu'il fait de l'actualité, se fond systématiquement dans celle du plus grand nombre, la vox populi. Elle se manifeste le plus souvent par l'ironie : les scientifiques sont de grands hommes, mais leurs découvertes sont bien piètres en regard des grands fléaux de l'humanité. Ainsi la thématique de la dérision peut-elle s'appliquer aux découvertes scientifiques et techniques, et par extension à leurs auteurs. » (Clavier, 2010 : p. 114)*

Nous indiquions que le repérage des marqueurs de positionnement éditorial dans la presse présentait des points communs avec les marqueurs de positionnement scientifique dans les écrits scientifiques (cf. *infra*) : dans les deux cas, ils révèlent des postures énonciatives. Formellement, les marqueurs se manifestent par un jeu de repères énonciatifs. Nous en donnions quelques exemples tirés de l'analyse des Causeries :

> *« Formellement, l'on passe du il (dans l'événement) au on doxique, au je (dans les commentaires), de l'assertion déclarative aux interrogatives et aux exclamatives, du mode indicatif au mode impératif. L'introduction de guillemets ne renvoie pas à des citations comme dans la presse contemporaine, mais à des commentaires qui sont des aphorismes, des proverbes, des dictons. La présence de traces repérables permet l'annotation des commentaires. » (ibid.)*

En conclusion, cette étude montre que l'organisation des connaissances s'appuie sur différents niveaux d'information pour créer des parcours thématiques dans des collections de presse numérisées anciennes. Après avoir constaté que la recherche d'événements, de dates, de lieux et de personnes, était un type de demandes répandu parmi les usagers, nous avons défini ce qu'est un thème dans la presse écrite. Contrairement à l'approche documentaire qui limite la notion de thèmes aux *sujets* présents dans un document, nous avons montré qu'un thème renvoie au positionnement éditorial d'un journal sur des événements, chaque événement traitant d'un ou plusieurs sujets. L'élaboration d'un parcours thématique nécessite trois étapes, les limites étant à mettre sur le compte du caractère automatisable des traitements.

- La première étape consiste à définir *a priori* des sujets dans une collection. Des méthodes d'extraction de connaissances linguistiques peuvent être appliquées à condition que les textes numérisés et océrisés soient corrigés.
- La seconde étape consiste à identifier des événements en lien avec les sujets et par numéro de journal. Les méthodes d'indexation et de classification automatiques de familles d'événements sont-elles actuellement opérationnelles ?
- La troisième étape consiste à repérer le positionnement éditorial sur un événement en tenant compte de la rubrique (quand elle existe) ou du genre de texte. Se pose ici la question de l'analyse automatique du métadiscours.

3.4.2. SCIENTEXT : les thèses de doctorat créent un « espace de sens qui favorise la recherche d'un *positionnement* scientifique »

Le titre de cette section paraphrase un article[174] écrit par Sylvie Dalbin et Brigitte Guyot (2007) sur les documents en action dans une organisation. Dans ce texte, les auteurs soutiennent que le document, en même temps qu'il crée un « *espace d'inscription* » (cf. le document comme *signe*), « *engrange des usages potentiels plus ou*

[174] « *Le document crée un espace de sens qui anticipe une action potentielle* ». (Dalbin et Guyot, 2007 : p. 60)

moins déterminés à l'avance » (*ibid.* p. 61) Les auteurs montrent que certains documents apportent des informations qui vont déclencher des actions au sein de l'organisation. Par exemple, un compte rendu de réunion anticipe diverses actions : lancer un projet, le financer, la répartition des tâches à assurer, etc. chacune de ces actions impliquant des services ou des acteurs différents. Les auteurs observent qu'il serait opportun de diffuser un compte rendu structuré par appariement d'informations et d'usages interrogeable pour une action précise, plutôt que de diffuser le document tout entier :

> *« Le diffuser tout entier, c'est suivre une orientation producteur-auteur, celui-ci se valorisant à travers cette production globale, alors qu'isoler les différentes unités d'information pour les articuler avec les espaces documentaires (les intégrer à une base de données), serait suivre une logique utilisateur associée à une action » (Dalbin et Guyot, 2007 : p. 61)*

De manière parallèle, nous souhaitions caractériser les contenus de thèses de doctorat en fonction de pratiques de recherche d'information observées (SCIENTEXT) :

> *« En ce qui nous concerne, nous étudions les pratiques de recherche d'information de manière située, considérant que l'activité principale de l'individu influence les attentes en matière d'information, les stratégies de recherche et plus largement l'activité informationnelle. Nous considérons ainsi que la conception d'un système nécessite que soient prises en compte les spécificités du contexte de la tâche de recherche d'information (Paganelli et al, 2002), ainsi que les caractéristiques textuelles et discursives des documents (Poudat et al, 2006). » (Clavier et Paganelli, p. 172-173)*

3.4.2.1. Sélectionner les informations à partir des pratiques observées

Les thèses de doctorat sont des documents structurés relativement longs, qui sont généralement « *consultés de manière fragmentée, non séquentielle et qui mettent en œuvre un grand nombre d'activités matérielles et cognitives (copier-coller, surlignage, annotations) qui sont autant de traces de l'activité informationnelle de l'individu (Hochon et al, 1994) (Mille, 2005)* » [175] (Clavier et Paganelli, 2012 : p. 173). Nous avions observé que les parcours de lecture n'étaient pas fondés sur des critères thématiques mais sur des critères privilégiant la voix, l'intention de l'auteur, sa

[175] A l'exception de la « lecture évaluation » qui suppose une lecture cursive et exhaustive des documents.

démarche scientifique, i.e son *positionnement.* L'accès au contenu d'un document suivant ce principe favorisait l'interprétation et la compréhension d'un document, il privilégiait des connaissances situées et permettait *in fine* au lecteur de positionner ses propres recherches (Clavier et Paganelli, 2010). Si nous adoptons le point de vue d'Yves Jeanneret sur les connaissances « *pour indiquer le travail productif des sujets sur eux-mêmes pour s'approprier des idées ou des connaissances* » (Jeanneret, 2000 : p. 119), alors les traces d'activités recueillies dans les passages annotés par les lecteurs et commentées par eux-mêmes renseignent sur cet effort d'appropriation :

> *« […] les annotations ajoutées par les lecteurs et les commentaires oraux associés à chaque passage de texte ont permis d'appréhender la manière dont le lecteur comprend le document qu'il consulte. Ces traces personnelles sont une manière pour lui de s'approprier le document et d'en interpréter le contenu (Mille, 2005). » (Clavier et Paganelli, 2012 : p. 174).*

Ainsi, l'analyse des discours (entretiens, traces d'activité commentées[176]) a révélé que les parcours de lecture différaient selon l'année de préparation de la thèse :

> *«[…] si les lecteurs cherchaient d'abord à « planter le paysage » (connaître les auteurs, les écoles de pensées, cerner la terminologie, etc.), ils aspiraient ensuite à se situer eux-mêmes (citer tel auteur plutôt que tel autre, s'inscrire dans un courant, adopter sa propre terminologie)… » (ibid.)*

Et que les thèmes avaient une importance toute relative contrairement à la recherche d'un positionnement scientifique :

> *« Ainsi, si les thèmes sont utiles pour sélectionner le document, puis les parties de thèses à consulter, ce sont les éléments métadiscursifs révélant le positionnement de l'auteur qui guident la lecture. »*

A l'issue de ces observations, nous avions d'abord procédé au typage des marques formelles, comme le soulignement et le surlignage par exemple, qui permettent de délimiter la taille des fragments :

> *« Nous avons analysé 158 fragments de textes. 129 d'entre eux comportaient des marques visuelles (soulignement, surlignage, etc.), 47 présentaient des annotations (notes, abréviations, mots-clés, symboles ; 148 ont été commentés oralement. Nous avons observé que les annotations et les commentaires permettaient d'identifier deux ensembles d'indices dans les*

[176] Cf. Méthode de constitution des corpus « informationnels » en 2.3.2.3.

> *fragments. Le premier ensemble opère au niveau du discours, l'autre au niveau de la textualité. » (Clavier et Paganelli, 2012 : p. 174)*

puis, nous avions indiqué leur position au sein de la structure des documents :

> *Autre caractéristique remarquable, la quasi-totalité des fragments annotés correspond à l'un des niveaux de structure du document suivants : partie entière (1), section entière (1), note de bas de page (13), sous-titre (10) et paragraphe (134), les quelque fragments restants correspondant à des « bouts de textes » disséminés dans la page. (Clavier et Paganelli, 2010)*

Cette analyse confirmait le paragraphe comme niveau d'information pertinent, même si tous les paragraphes n'étaient pas complets :

> *« Toutefois, les paragraphes ne sont pas toujours complets : sur 134, 83 sont tronqués. On observe que les fragments tronqués renvoient à des unités de langue, telle que la phrase. Il y a davantage de phrases complètes (257) que de phrases incomplètes (87). Enfin, dans ce jeu de poupées russes, les phrases incomplètes correspondent à leur tour à des unités syntagmatiques du rang de la proposition (46), puis du groupe nominal (24). Les fragments restants renvoyant à diverses catégories marginales (entités nommées, références bibliographiques isolées, etc.). » (Clavier et Paganelli, 2010)*

… et que les phrases se répartissaient quasiment pour un tiers en début, en milieu et en fin de paragraphe. Nous recherchions alors les « fameux » marqueurs de positionnement sur lesquels travaillaient nos collègues du projet SCIENTEXT (Rinck et al, 2007).

3.4.2.2. Identifier des marqueurs de *positionnement*

Dans le projet de recherche SCIENTEXT déposé auprès de l'Agence Nationale de la Recherche en 2007, la définition de la notion de *positionnement* est donnée. Caractéristique de l'écriture scientifique, le *positionnement* se manifeste dans les procédés linguistiques qui révèlent « *la singularité d'un auteur, son apport spécifique – la justification de sa démarche scientifique – et le raisonnement de l'auteur, ce sur quoi il s'appuie, les preuves qu'il emploie, les relations logiques qu'il établit – la qualité de l'analyse scientifique* »[177] Le *positionnement* n'est pas une catégorie linguistique : il se

177 Scientext : *un corpus et des outils pour étudier le positionnement et le raisonnement de l'auteur dans les écrits scientifiques*. (dir. F. Grossmann et A. Tutin), document final soumis à l'ANR en 2007.

réalise dans un ensemble de marqueurs émanant de plusieurs niveaux d'analyse (lexique, syntaxe, phraséologie) (Tutin, 2007). L'état de l'art en linguistique indique que les marqueurs de positionnement relèvent de l'épistémicité et de l'évidentialité. Pour Eva Vold (2008), la modalité épistémique « *concerne nos connaissances du monde, elle exprime les jugements du locuteur par rapport à la fiabilité du contenu propositionnel.* ». Quant aux évidentiels, ils renvoient *« à l'expression du mode de création et/ou de récolte de l'information »* (Dendale et Tasmowski 1994 : p. 4). Selon ces auteurs, Ils apparaissent dans trois sources : la perception (visuelle ou auditive), l'emprunt et l'inférence. Nous donnions des exemples mentionnés par les collègues impliqués dans le projet SCIENTEXT :

> *« A l'écrit, ils s'actualisent dans les verbes de perception (on constate que), les tournures impersonnelles (il semble évident que, il apparaît que), les différents modes de désignations des références à autrui (la citation d'auteur, la référence bibliographique, les noms de courants scientifiques) qui contribuent au cadrage historique et conceptuel des notions. » (Clavier et Paganelli, 2010).*

A la suite d'auteurs spécialistes des discours scientifiques, nous indiquions que ces marqueurs étaient caractéristiques d'une posture dite de « surénonciation » caractéristique de ce type d'écrits :

> *« Selon Jacobi (1999 : 15), l'une des caractéristiques majeures des discours de communication scientifique serait surtout de permettre de consolider la place et la légitimité de leurs auteurs. La thèse de doctorat manifeste cette propension en raison du dispositif d'intronisation qui entoure l'auteur pour être reconnu par la communauté scientifique. L'écriture de la thèse consiste alors à mobiliser des connaissances, à faire valoir son point de vue, à construire son identité de chercheur. Les procédés linguistiques qui concourent à établir cette posture « surplombante » relèvent de la surénonciation. » (Clavier et Paganelli, 2010)*

D'un point de vue énonciatif, les écrits scientifiques ont la particularité d'être des énoncés non embrayés, ce qui leur confère un caractère objectif, plus empreint de scientificité. L'analyse polyphonique se réalise donc dans un contexte *d'effacement énonciatif* (Rabatel, 2004), dans lequel le locuteur s'efface au profit d'un énonciateur universel (le *nous* de majesté, le *on* à valeur doxique, les tournures impersonnelles) et c'est donc par l'introduction du sujet modal dans l'énoncé que se définit la prise en charge de l'énonciation. Les marqueurs d'épistémicité et d'évidentialité mentionnés *supra* contribuent d'une part à situer l'approche de l'auteur par rapport à celles des

autres (se positionner), d'autre part, à justifier ces points de vue (argumenter, prouver).

Revenons aux fragments. Nous relevions deux ensembles de marqueurs. Le premier ensemble opérait au niveau du discours :

> *Dans le premier cas, les indices recueillis sont des évaluatifs, des axiologiques et des catégories relevant de l'épistémicité et de l'évidentialité. Nous avons ainsi retrouvé les marques linguistiques mentionnées dans (Dendale, 1994) (Boch et al, 2007) (Rinck, 2007), même si les catégories sont en moindre quantité que dans les études mentionnées (ibid.) (Clavier et Paganelli, 2012 : p. 175)*

l'autre, au niveau de la textualité :

> *Dans le second cas, les indices recueillis permettent de localiser des énonciations suivant leur position dans le document. Nous suivons ainsi A. Berrendonner (1997) pour qui il existe des « pointeurs métadiscursifs » qui sont des déictiques (ici, ci-contre), des extraits de textes (dans la première partie) ou encore des localisations floues (dans ce passage) et pour qui le document est un « espace textuel vectorisé »*[178].

Pour éviter toute confusion entre les deux niveaux de marqueurs, nous choisissions de parler de marqueurs métadiscursifs lorsqu'ils permettent de se repérer sur le plan cognitif et de marqueurs métatextuels lorsqu'ils permettent de se repérer dans le document.

3.4.2.3. Interpréter et représenter les marqueurs de *positionnement*

A cette étape du mémoire se pose la question de l'organisation des connaissances. La méthodologie mise en place, visant à identifier les pratiques informationnelles, à recueillir les traces d'activité et à analyser le contenu prépare la tâche d'indexation. Le recueil des connaissances n'est donc ni au « ras-de-la langue » comme dans le cas de l'indexation automatique, ni au « ras des pratiques » comme dans le cas de l'indexation collaborative. Le recueil des marqueurs de positionnement consistait à dresser des listes de marqueurs métadiscursifs et métatextuels. Mais cette tâche

178. Pour Alain Berrendonner, le texte est « *l'ensemble ordonné des énonciations accomplies successivement au fil d'un discours* » et « *l'espace vectorisé* » une « *schématisation du texte comme espace.* » (Berrendonner, 1997 : p. 221).

présentait des difficultés relatives au jugement de catégorisation et à leur repérage au sein des textes :

> *« L'étude des indices de positionnement dans le corpus (Clavier, 2010) a révélé deux limites importantes : d'une part, une grande diversité de marqueurs renvoyant à de nombreuses catégories sémantiques qui se recoupent parfois, ou qui sont difficiles à cerner. D'autre part, une dissémination des indices dans tout le texte, ce qui rend caduque toute tentative d'indexation suivant cette approche. » (Clavier et Paganelli, 2012 : p. 176)*

Cette situation s'explique par deux raisons : d'une part les marqueurs qui relèvent du discours sont des segments de rang et de nature différents, non linéaires mais quand même « accrochés à la structure textuelle » (Péry-Woodley et Scott, 2006) ; d'autre part, l'interprétation des marqueurs ne peut se faire indépendamment les uns des autres, puisque leur caractéristique est « d'instruire une relation » :

> *« Ces marqueurs langagiers ont comme caractéristique d'être discontinus, de relever du discours ou du métadiscours et, pour (Ho-Dac et al, 2008), ils ne doivent pas être confondus avec des marqueurs de segmentation mais sont plutôt des indices qui « participent à instruire une relation de continuité ou de discontinuité entre deux segments » (ibid.)*

Nous illustrions cette particularité par un exemple en lien avec la terminologie :

> *« Par exemple, si les sujets cherchent à « comprendre les enjeux scientifiques de la terminologie » (Annexe A), ce n'est pas la terminologie décontextualisée qui les intéresse, mais bien « la terminologie d'un auteur », « l'univers d'un auteur ». Il faut donc décrire la nature de la relation qui unit la terminologie à son auteur. » (Clavier et Paganelli, 2010)*

En cherchant à établir des relations entre les connaissances propres au positionnement scientifique, nous posions les bases d'un système d'organisation combinant des connaissances indicielles et non indicielles, c'est-à-dire relationnelles. Nous proposions d'en rendre compte dans une grammaire à trois positions qui s'applique à la phrase, ou éventuellement au paragraphe. Nous indiquions que ces catégories définissent « le périmètre triangulaire » du positionnement (Clavier et Paganelli, 2012) :

1. Les expressions qui renvoient à un jugement ou une appréciation subjective de l'auteur (l'adhésion, l'atténuation, la critique, le consensus, *etc.*)

2. Les expressions qui dénomment un thème (des termes, des concepts, des contenus propositionnels, *etc*.).
3. Les expressions qui mentionnent dans quel environnement (ou quel repère) on se situe : ce peut être dans le discours (dates, lieux, références à autrui *etc*.) ou dans le document (chapitre, partie, *etc*.). Voici des exemples comportant des indices de positionnement. Ces énoncés sont extraits de la thèse lue par l'un des sujets interrogés dans (Clavier, 2010).

Nous donnions des exemples d'énoncés extraits de l'une des thèses :

E1. Il demeure cependant indéniable que l'hypertexte est un terme qui fait aujourd'hui partie de notre culture commune. (Ertzscheid, sujet 3)

E2. Sans point commun apparent avec l'idée de Nelson, il est intéressant de remarquer comment, au point actuel de l'évolution technologique, les deux définitions entrent sans peine en résonance, laissant entrevoir un champ épistémologique à la fois ouvert et complexe dans lequel les associations de l'un font écho aux « dérives » de l'autre. (Ertzscheid, sujet 3)

E3. Nous défendons dans ce travail la thèse selon laquelle l'hypertexte n'est pas un épiphénomène de nature informatique assimilable ou réductible à l'un des sphères de la réalité qui l'emploie. (Ertzscheid, sujet 3)

Nous indiquions que chaque énoncé comportait les trois catégories de marqueurs :

1. Les expressions qui signalent un positionnement (affirmation, constat, thèse) : *il demeure indéniable que* (E1), *il est intéressant de remarquer que* (E2), *nous défendons la thèse selon laquelle* (E3).
2. Les expressions qui décrivent un thème : *hypertexte* (E1, E2, E3).
3. Les expressions qui permettent de localiser un point de vue, (temps, lieu ou angle d'approche) : *aujourd'hui* (E1*), sans point commun avec l'idée de Nelson* (E2) *En prenant l'angle critique qu'offre l'analyse des hypertextes littéraires* (E3) (document, chapitre, localisation floue) : *dans ce travail* (E3).

Nous précisions que cette grammaire était très contrainte, et qu'il serait possible d'en relâcher les contraintes :

> *« Bien que n'ayant pas encore de données précises sur le rendement de ce modèle, nous avons fait le choix (dans un premier temps) de proposer le modèle le plus contraint puisqu'il nécessite la présence simultanée de trois catégories de connaissances. Le positionnement a donc un statut composite : il combine des éléments langagiers de nature indicielle et relationnelle et renvoie à des marqueurs de différents niveaux (lexical, syntaxique). » (Clavier et Paganelli, 2012 : p. 177)*

3.4.2.4. Bilan : des réponses partielles sur des aspects théoriques de l'organisation des connaissances

La proposition que nous avons faite soulève deux questions, l'une relative au statut théorique des connaissances que nous manipulons et l'autre relative aux méthodes d'extraction automatique pour les identifier, les coder et les formaliser.

En ce qui concerne le premier point, quelle est la nature des connaissances manipulées ? Pratiquement toutes les disciplines qui relèvent des humanités se posent ces questions, en premier chef la philosophie, si bien qu'une part importante des travaux en SIC puise ses référents théoriques en philosophie. D'autres théories cognitives ont bénéficié d'une extraordinaire longévité, comme par exemple la théorie du prototype d'Eleanor Rosch en 1973 (*Natural Categories*). Michèle Hudon et Widad Mustapha El Hadi s'interrogent cependant sur l'« absolue nécessité » de recourir à une théorie unifiée de l'organisation des connaissances et s'il elle ne reposerait pas plutôt sur un ensemble disparate d'éléments théoriques empruntés à d'autres disciplines. Elles se demandent s'il est essentiel de normaliser la terminologie de l'organisation des connaissances. Ou encore, par quoi remplacer le modèle traditionnel d'organisation à base disciplinaire (Hudon et Mustapha El Hadi, 2013 : p. 11). L'étude que nous avons réalisée donne des réponses partielles à certaines de ces questions.

Une première réponse réside dans la place toute relative de la terminologie pour organiser les connaissances. Nous avons constaté que le positionnement de l'auteur était une notion motrice pour guider l'activité de consultation de thèses de doctorat et que la terminologie jouait un rôle secondaire. Nous signifions par là que la terminologie d'un domaine n'intéresse pas les lecteurs en tant que telle, mais la

terminologie d'un auteur, en lien avec son raisonnement scientifique et la construction de sa pensée. D'où la question suivante : cette observation est-elle généralisable à d'autres disciplines que les SHS, souvent marquées par une culture de la pensée individuelle et par l'affirmation d'un positionnement singulier ? à d'autres types d'écrits, comme les articles de revue par exemple ?

Une deuxième réponse réside dans l'importance des discours pour organiser les connaissances. La recherche d'information de type professionnel cherche des réponses précises, en réponse à des demandes finalisées et orientées vers certaines tâches. Cette assertion ne signifie pas que les lecteurs recherchent des informations factuelles et décontextualisées. Nous considérons l'extraction terminologique comme relevant d'une perspective décontextualisée. Or, d'après François Rastier, la propension à se focaliser sur les termes et les concepts en science est à mettre sur le compte d'une méfiance vis-à-vis des langues naturelles et des normes sociales d'écriture :

> *« La question du texte scientifique reste difficile à poser : la tradition scientifique occidentale fait de la science une affaire de concepts et de termes, non de textes, car elle tient que l'objectivité est indépendante de la différence des langues et des normes textuelles. » (Rastier, 2005)*

Et, nous rajoutons que pour la recherche d'information, la question des langages (d'indexation versus d'interrogation) constitue également le problème central de l'accès à l'information. Si bien qu'il y aurait deux raisons différentes, culturelle et technique, pour lesquelles les dispositifs informationnels accorderaient une place plus importante aux termes et aux concepts qu'aux discours. Or, la mise en discours de la terminologie, n'est-elle pas la condition *sine qua non* d'une appropriation réussie des savoirs ainsi que nous le constations dans un environnement pédagogique (Clavier et Lafont-Terranova, 2005) ?

Une troisième réponse réside dans la proposition d'un modèle de représentation de l'espace et du temps pour organiser les connaissances. L'expérience acquise pour modéliser des faits sémantiques en morphologie dérivationnelle (cf. la première partie de ce document, 1.2.3.3.) nous avait enseigné

que la formalisation doit s'appuyer sur des catégories générales distinctes de l'objet à décrire, en l'occurrence ne pas utiliser les langues naturelles pour formaliser les langues naturelles. Nous avons fait le choix d'un modèle localiste qui fait partie des références classiques en TAL pour formaliser les connaissances[179] et qui présente des liens avec le *hic et nunc* en linguistique de l'énonciation rappelant la nécessaire prise en compte du contexte pour analyser les discours. Cette proposition est développée dans la publication de 2010, et comporte quelques aménagements en 2012. Il serait opportun d'approfondir ce modèle, qui comporte en l'état actuel des imprécisions, voire même des incohérences. Le cadre localiste permet de justifier la grammaire à trois positions présentées supra, et permet de décrire le lien entre un thème et les relations qui se nouent avec le document (le métatexte) ou avec le discours (le métadiscours). Le cadre localiste comporte trois entités :

- un repère : les expressions indicielles qui décrivent un thème ;
- un environnement spatio-temporel : les expressions qui permettent de localiser un point de vue (une date, un lieu, un courant de pensée) ou de localiser un passage dans le document (une page, un chapitre, une annexe) ;
- une relation : les expressions qui signalent un point de vue de l'auteur.

Nous justifions le choix de ce modèle sur le plan cognitif :

> *« Sur le plan cognitif, l'annotation de textes scientifiques révèle une partie des activités liées à la lecture savante, et permet d'établir un lien avec le support et la textualité. A ce niveau, il nous semble possible de parler du positionnement comme d'une posture qui serait le fil d'Ariane entre la lecture, l'annotation et l'écriture. Comme mentionné supra, l'exercice d'écriture scientifique induit la recherche d'un positionnement, l'objectif étant d'acquérir un point de vue « surplombant ». (Clavier et Paganelli, 2010)*

et sur le plan linguistique :

[179] Voir pour un état de l'art sur les hypothèses localistes (Lyons J. CH 6.7 Le localisme in *Sémantique linguistique*, Larousse, 1990, 388sqq ; Rousseau A. « Espace, référence, représentation. Réflexions sur quelques conceptualisations de l'espace », *Faits de Langue, 1/1993*, pp. 151-162 ; Desclés J.-P. « La prédication opérée par les langues (ou à propos de l'interaction entre langage et perception) », *Langages, 103, 1991*, Larousse, pp. 83-95.

> *« Sur le plan linguistique, l'hypothèse localiste dans sa version cognitive pourrait offrir un cadre de description satisfaisant puisqu'elle permet de penser le positionnement par le biais d'interactions entre des représentations de différents niveaux, le langage étant l'un de ces niveaux. La question de l'ancrage catégoriel se pose alors en des termes métadiscursifs et non pas propositionnels. Cela signifie que contrairement aux versions fortes du localisme dans lesquelles seules les catégories spatio-temporelles figurent dans certains schémas de la prédication (les prépositions, le temps grammatical, l'aspect, les cas, les adverbes etc.), une version métadiscursive du localisme permettrait de décrire les interactions entre un auteur et son destinataire dans une même communauté, et les relations internes au discours (Tutin, 2007). » (ibid.)*

C'est à ce niveau de description et de caractérisation des informations que nous avons utilisé les connaissances sur la langue et le discours pour donner du sens, entendons par là, donner une cohérence aux traces d'usages.

3.5. Bilan critique

Les travaux qui relèvent de cette seconde partie présente une évolution majeure de notre parcours qui s'est marquée par la prise en compte du contexte pour recueillir des connaissances. Les questionnements relatifs à l'organisation des connaissances se sont posés à nous progressivement, nos premiers travaux conduits au GRESEC étant *a priori* éloignés de cet objet d'étude, puisqu'ils portaient sur l'analyse d'usages et de pratiques informationnelles en milieu médical. Cependant, avec le recul, nous comprenons plus précisément comment les systèmes d'organisation des connaissances peuvent avoir une influence déterminante sur les pratiques des usagers. Questionner les pratiques consiste aussi à interroger les « *pratiques informationnelles préexistantes* » (Le Marec, 1997 : p. 540), cette dimension permettant de se dégager du *hic et nunc* et de l'infinité des usages individuels observés, et d'accorder une place importante à la culture informationnelle. Inversement, nous observions que les activités des usagers aux prises avec l'information, révélaient des éléments de connaissance qui ne sont pas du tout pris en compte dans les dispositifs, contraignant ainsi les individus à développer des stratégies de contournement pour accéder à l'information. **En fin de compte, notre travail cherche à répondre à deux questions : comment identifier les éléments**

d'information propres au contexte d'usage et de pratiques ? Comment ces éléments d'information contribuent-ils à l'organisation des connaissances ?

Concernant le premier point, le principal apport de nos travaux est d'avoir développé une méthodologie de constitution de corpus destinée à « capter » le contexte. Le contexte est une entité objective et subjective, ainsi que le montrent les approches contextualisées de la recherche d'information. Nous avons mis en œuvre trois types de corpus qui se distinguent par la nature des informations prises en compte pour recueillir les données : les corpus linguistiques qui comportent des formes langagières ; les corpus documentaires qui se présentent comme des sources d'information pour des usages spécifiques ; les corpus « informationnels » qui recueillent des traces d'activités, des contenus, des commentaires sur les pratiques. Les trois catégories de corpus prennent tous en compte des critères sociaux.

Ainsi, dans les corpus linguistiques, une forme langagière comme *cancer* est-elle une unité lexicale propice à la connotation : la métaphore. Dans les discours, elle donne lieu à des représentations négatives de la maladie. Les métaphores sur le *cancer* étant figées et véhiculant depuis des siècles les mêmes stéréotypes, elles favorisent des représentations sociales partagées qui se manifestent par exemple, sous forme de dénominations. Dans ce cas, le contexte est une entité objective : il désigne le co-texte. Ce dernier met en évidence le lien entre la mémoire sociale, les discours, et « l'existence historique » des énoncés, perspective qui est connue sous l'expression « mémoire discursive » utilisée par Jean-Jacques Courtine (1981) pour indiquer que « *le langage constitue la matière, ici tissu de la mémoire* » (Paveau, 2006 : p. 91).

Dans les corpus documentaires, les ressources retenues sont des documents spécialisés obéissant à des normes d'écriture : c'est le cas des dictionnaires encyclopédiques, des terminologies, des glossaires et des articles scientifiques. Mais les ressources sont aussi des dispositifs techniques, tels les forums de santé qui comportent des informations médicales « vécues ». Dans les deux cas, le contexte est une entité objective. Soit le contexte se traduit par des normes sociales et scientifiques qui contraignent la production éditoriale, les usages des documents. Soit

il renvoie à la situation vécue par l'usager, la maladie réelle ou redoutée, et induit des types d'informations spécifiques comme les témoignages empreints d'émotion.

Dans les corpus « informationnels », nous avons recueilli les traces d'activités informationnelles commentées par les sujets. Ces traces et ces discours révèlent les effets des contextes sur les significations. C'est ainsi que nous sommes parvenue à hiérarchiser et interpréter des éléments du contexte qui ont une influence sur la recherche d'information. Par exemple, le positionnement scientifique est une posture recherchée pour se faire accepter des pairs. Lors de la recherche d'information, cet enjeu se traduit par des stratégies informationnelles spécifiques et par des marqueurs idoines. La recherche de la terminologie apparaît comme un élément de moindre importance pour les chercheurs en SIC que pour les médecins spécialistes. Ainsi, notre méthode met à jour divers éléments du contexte qui comprend *« des acteurs, l'activité principale que ces acteurs mènent dans le cadre de leur travail, un environnement informationnel constitué de documents et de dispositifs, un environnement socio-organisationnel »* (Paganelli, 2012 : p. 129-130). C'est cette articulation entre différents éléments qui est individuelle, et d'une certaine manière subjective.

C'est la raison pour laquelle nous considérons que les conditions de recueil de l'information s'inscrivent dans une perspective de médiation et non de normalisation de l'information. Précisons à ce niveau qu'il existe une méthode alternative, beaucoup plus répandue que le recueil de traces d'activités pour sélectionner les informations. Il s'agit de s'appuyer sur des enquêtes d'usage existantes ou proches de l'information à caractériser. Cette méthode a été utilisée dans le projet CaNu, les enquêtes indiquant que la recherche de thèmes et d'événements est plébiscitée par le grand public. Cette méthode est beaucoup moins riche que celle utilisée pour l'analyse du positionnement de l'auteur, car elle ne dit pas précisément ce que les usagers entendent par événements. On en vient donc à traiter les événements comme des catégories objectives et non subjectives. Or, un historien, un sociologue des médias, un scientifique et les profanes appréhendent-ils les événements de la même façon ? Nous n'en sommes pas du tout certaine.

Concernant le second point, relatif à l'utilisation des informations pour l'organisation des connaissances, notre principal apport réside dans la mise en évidence de niveaux de structuration de l'information. Nous avons toujours travaillé les contenus en prenant soin de distinguer les unités de langue (unités lexicales simples ou complexes, la proposition et la phrase), les niveaux de la textualité (la structure logique du texte, le paragraphe), les paliers supérieurs au texte (le paratexte). En outre, l'étude sur le positionnement de l'auteur cherche à « accrocher » le métadiscours à la textualité.

A l'exception des forums de santé, nos études s'inscrivent dans le champ de l'information spécialisée. Toutes mettent en avant le rôle primordial des documents pour les sujets interrogés. Les médecins spécialistes privilégient les articles scientifiques qui sont évalués par des revues reconnues internationalement ; les spécialistes de littérature s'intéressent aux œuvres, qu'elles soient éditées dans des livres ou ébauchées dans les manuscrits ; les doctorants s'intéressent aux thèses qu'ils conçoivent comme modèle d'écriture et comme source d'information. Les informations relatives aux documents, telles les métadonnées bibliographiques ont été intégrées dans nos analyses de corpus. Par exemple, l'étude sur l'information professionnelle tient compte de l'évolution d'une notion au fil des années de publication des revues ; elle considère le statut des auteurs, les mots-clés qui caractérisent leur recherche, la discipline de la revue. Mais les documents consultés ont aussi été analysés comme supports de traces d'activités humaines. Ces dernières, confrontées aux textes et aux commentaires oraux, constituent un complexe d'observation des usages et des pratiques informationnelles.

Les techniques utilisées pour analyser l'information sont au nombre de trois. Dans le corpus « Information professionnelle », nous avons procédé à des analyses de contenu destinées à catégoriser les contextes de trois occurrences : l'information professionnelle, l'information spécialisée et l'information scientifique et technique. Dans le projet CaNu, nous avons recherché les contenus qui traitent de sciences et de techniques, recueillant ainsi des textes issus des *Causeries* ainsi que des gravures. Nous avons utilisé plusieurs méthodes d'analyses de discours. Dans le *Progrès Illustré* (projet CaNu), nous avons cherché à articuler les événements aux thèmes et avons

considéré ces derniers comme une manifestation de l'angle journaliste. Rompant ainsi avec une conception référentielle du thème, nous empruntions nos références théoriques aux spécialistes des discours de presse (Ringoot et Robert-Demontrond, 2004). Dans l'étude sur le positionnement scientifique en revanche, nous privilégions une analyse des discours qui s'appuie sur l'organisation textuelle au sens des travaux de linguistique réalisés par les collègues du projet SCIENTEXT ou (Ho-Dac et al. 2012). Dans l'étude sur le *cancer* enfin, nous avons recouru aux méthodes statistiques de cooccurrences.

Se pose enfin la question de l'intérêt d'une telle approche pour l'organisation des connaissances. Nous offrons deux types de réponses : l'une appliquée, l'autre plus théorique. En premier lieu, l'orientation appliquée s'est essentiellement posée dans les études qui considéraient que les méthodes d'indexation automatique avaient un rôle important à jouer (CaNu et SCIENTEXT). Cet objectif revenait à appréhender l'organisation des connaissances comme un produit, les index. Nos propositions ne vont pas jusqu'à la mise en œuvre concrète de schémas d'organisation, contrairement à la première partie de ce mémoire. Ce que nous proposons en revanche, ce sont des axes d'organisation des connaissances identifiés à partir de plans de structuration de l'information, de catégories de documents et de pratiques informationnelles. Ainsi, après avoir constaté que le positionnement était identifié par des catégories linguistiques, qu'il avait une pertinence pour consulter des documents scientifiques et que cette pertinence était validée expérimentalement, il devenait possible de revenir sur les méthodes d'extraction automatiques. C'est ainsi qu'une thèse a été proposé par Agnès Tutin sur ce sujet[180]. Nous constatons *in fine* qu'envisager l'organisation des connaissances sous l'angle des applications nécessite des collaborations étroites avec des partenaires scientifiques de disciplines comme le TAL et/ou l'ingénierie des connaissances.

[180] Hatier Sylvain (en préparation), *Extraction et catégorisation de lexiques transdisciplinaires d'articles scientifiques de sciences humaines en vue de l'indexation automatique* – Université Stendhal Grenoble 3, LIDILEM.

En second lieu, les questions relatives à l'organisation des connaissances se sont posées sur le plan théorique et méthodologique. Nous avons utilisé diverses méthodes pour hiérarchiser et structurer l'information tout en intégrant les enquêtes d'usages et de pratiques. Ce faisant, nous posions la question de la nature des connaissances mises en évidence, de leur mode d'organisation et de leur dénomination. Ces dernières sont éloignées d'une conception philosophique, linguistique ou cognitive des connaissances. Intimement liées aux pratiques sociales, à la place du document dans les pratiques informationnelles, les connaissances sont marquées par des enjeux professionnels individuels et collectifs et s'éloignent d'une conception du savoir disciplinaire... Repérées par des analyses micro-sociales, nous posons cependant la question de leur portée.

Conclusion

Au terme de cette présentation, nous dressons un bilan de nos objets de recherche et des perspectives dans lesquelles ils ont été abordés. Nous proposons ensuite des orientations pour notre activité de recherche future, certaines pistes nous paraissant opportunes.

L'organisation des connaissances, en tant qu'objet d'étude, a été analysée dans le cadre de la recherche d'information, un domaine de recherche investi par plusieurs disciplines. Nous avons tout d'abord abordé la recherche d'information dans une perspective technique orientée vers la conception de systèmes en laboratoire. Les données textuelles constituent alors un réservoir de connaissances susceptibles d'être représentées, codifiées et formalisées par des technologies du langage. Nous avons ensuite abordé la recherche d'information dans une perspective sociale orientée vers l'observation d'usages et de pratiques informationnelles d'individus, essentiellement en situation professionnelle. Les discours recueillis à partir d'entretiens d'usagers en situation de recherche d'information, de documents consultés, de dispositifs utilisés, etc. constituent alors un réservoir d'informations susceptibles d'être représentées sous formes de connaissances. Ces orientations sont complémentaires à plus d'un titre.

Sur un plan technique qui contribue au développement d'interfaces de recherche, nous avons exploré des méthodes de TAL favorisant une représentation des connaissances fondée soit sur des niveaux de langue soit sur des niveaux de la textualité, comme le genre textuel. Ces modes de représentations sont utiles pour améliorer l'indexation en texte intégral et la classification automatique. Cependant, l'observation de pratiques informationnelles a montré que la recherche plein-texte était souvent insuffisante. Aussi, il paraît utile de développer d'autres formes d'accès tels des parcours thématiques pour accéder à des collections patrimoniales, ou des parcours qui privilégient la voix de l'auteur pour consulter des documents scientifiques. La diversification des modes d'accès à l'information et plus largement des formes de médiation des dispositifs informationnels n'est pas l'apanage des sciences de l'information et de la communication. Par exemple, dans le projet Métilde

qui se donne pour objectif de valoriser le fonds numérisés et transcrits des manuscrits de Stendhal, nos partenaires, respectivement informaticien et spécialiste de littérature, recourent-ils à ce type de proposition. La différence cependant, réside dans la perspective envisagée. En TAL et en littérature, le choix des parcours s'appuie sur des usages supposés ou souhaités. En SIC, les études d'usages (ou de non-usages) et de pratiques sont réellement constatées. Les propositions de parcours qui font suite à ces études, sont pensées pour être en adéquation avec les pratiques observées. Cependant, dans un cas comme dans l'autre, les usages ne se décrètent pas. Marie Després-Lonnet, qui reprend les propos de Bernard Miège (1997), indique au sujet d'Europeana, qu'il y a une confusion entre la capacité technique de « rendre accessible » des données par leur « mise en ligne » et « l'accès aux savoirs », c'est-à-dire, « la possibilité effective donnée à un individu de s'approprier de nouvelles connaissances » (Miège, 1997, cité par Després-Lonnet, 2009 : 19). Par conséquent, au sujet de Métilde, deux cas de figure se présentent pour « élargir les publics ». S'il s'agit du grand-public, il convient de mettre en place des dispositifs de médiation humaine pour amener les non-usagers à venir sur la plate-forme des manuscrits en ligne. S'il s'agit de publics spécialisés, il y a tout intérêt à développer des plates-formes qui s'intègrent au mieux dans leurs pratiques informationnelles. Dans les deux cas, aucune garantie n'est donnée sur le fait qu'ils deviendront des usagers assidus.

Sur un plan théorique et méthodologique, notre contribution s'inscrit dans un ensemble de travaux sur l'organisation des connaissances relevés par Birger Hjørland (2008) ainsi que par d'autres chercheurs (Broughton et al. 2005 : p. 141). Ces auteurs dressent un état des lieux de l'organisation des connaissances et identifient sept approches différentes [181] :

1. Les systèmes de classification utilisés dans les bibbliothèques et les bases de données, comprenant la classification décimale de Dewey et la classification décimale universelle (depuis la fin du 19ème siècle). Les unités de référence sont les documents, l'organisation des connaissances repose sur le principe de la classification de savoirs scientifiques et disciplinaires ;

[181] Nous résumons les propos, sans les traduire fidèlement.

2. La classification analytique par facettes (Ranghanathan, dans les années 1930). Le principe d'organisation repose sur des idées fondées sur « l'intuition rationnelle » ;
3. La tradition de l'*information retrieval* (depuis les années 1950). Les unités sont des mots, des relations de cooccurrences entre les mots issus de documents. L'unité de référence est l'information.
4. La perspective orientée *usage* (depuis les années 1970). Les unités sont des structures individuelles, cognitives.
5. L'optique bibliométrique (depuis les années 1963) Les unités considérées sont les modes de citations entre les documents.
6. L'approche analytique par domaine (depuis les années 1995). Les unités ne sont pas des connaissances éternelles mais des savoirs organisés en domaines appréhendés dans les communautés de discours résultant de la division du travail dans une société.
7. D'autres perspectives plus récentes comme les approches sémiotiques critiques herméneutiques, les analyses de discours et les approches fondées sur les genres (naturels), ainsi que les travaux sur les représentations des documents, les langages de balisage, l'architecture des documents.

Nos recherches sur l'organisation des connaissances se situent dans les perspectives 3) et 7) qui mettent l'accent sur les représentations. Ces dernières sont le résultat de traitements automatiques appliqués à des textes, de représentations sociales appréhendées dans les discours ou de standards et normes utilisés pour décrire des documents. Nos études présentent également un lien avec la perspective 6) dans la mesure où la recherche d'information est considérée comme une activité sociale « au-service » d'autres activités, que ce soient des tâches, des actions, des discours, etc. La recherche d'information apparaît alors comme un liant social. Toutefois, nous ne pensons pas que l'approche par domaines soit la plus adéquate pour travailler les connaissances. Pour Birger Hjørland, cette approche est une alternative à la perspective cognitive : elle suppose un point de vue sociologique[182] ce qui nous

[182] *« Domain analysis is a sociological-epistemological standpoint. The indexing of a given document should reflect the needs of a given group of users or a given ideal purpose. In other words, any description or representation of a given document is more or less suited to the fulfillment of certain tasks. A description is never objective or neutral,*

semble très pertinent. En revanche, les domaines présentent l'inconvénient de ne solliciter qu'une seule dimension de la textualité : les lexiques de spécialité. Les genres sont laissés de côté, alors qu'il a été montré que la portée typologisante des genres était utilisée à des fins classificatoires.

En guise de perspective, nous formons le projet de développer notre activité de recherche dans trois directions.

Tout d'abord, en lien avec le programme Métilde dont le terme est prévu en 2014, nous comptons poursuivre nos recherches sur la place du patrimoine numérique dans les activités professionnelles des individus. Cette orientation définie avec Céline Paganelli correspond au programme de recherche mis en place pour le contrat quinquennal 2011-2016 (cf. 1.2.4 *Recherche d'information : modélisation et usages*), et s'inscrit dans la continuité de l'ouvrage que nous avons co-dirigé sur l'information professionnelle en 2013. Alors que le patrimoine numérique est généralement considéré comme un objet culturel « grand public », nous souhaitons nous concentrer sur les publics de spécialistes, les archives patrimoniales pouvant être appréhendées comme des ressources spécialisées. Ce faisant, elles pourraient bénéficier de l'attention d'acteurs traditionnellement investis dans les politiques de traitement et de diffusion de l'information spécialisée.

Ensuite, nous souhaitons approfondir la notion d'événements dans la perspective d'une indexation de contenus médiatiques. Ce programme de recherche pourrait trouver un accueil favorable auprès des membres du projet Cogis[183] qui mêle des spécialistes des médias et des scientifiques spécialistes des catastrophes naturelles[184]. Il s'agirait de caractériser les discours des scientifiques et des médias,

and the goal is not to standardize descriptions or make one description once and for all for different target groups » (Hjørland, 2008 : p. 16)

[183] Projet CoGIS (Communication, Inondations, séismes), coordonné par Benoit Lafon, impliquant trois laboratoires des Universités Stendhal et Joseph Fourrier : le GRESEC, le LTHE (Hydrométéorologie), ISTerre (Sismologie) – dans le cadre des Projets d'Exploration Premier Soutien (Peps) financé par le CNRS, 2012-2013.

[184] Projet CoGIS (Communication, Inondations, séismes), coordonné par Benoit Lafon, impliquant trois laboratoires des Universités Stendhal et Joseph Fourrier : le GRESEC, le LTHE (Hydrométéorologie), ISTerre (Sismologie) – dans le cadre des Projets d'Exploration Premier Soutien (Peps) financé par le CNRS, 2012-2013.

de cerner les propriétés discursives des événements dans une perspective d'indexation tenant compte de la polyphonie des sources énonciatives convoquées dans les médias.

Enfin, à plus long terme, nous aimerions participer à la mise en place d'un projet de recherche sur la santé visant à étudier l'influence des mutations récentes de l'information médicale sur les pratiques des acteurs du domaine. Nous nous intéresserions plus spécifiquement aux pratiques informationnelles des institutions de veille sanitaire, aux dispositifs d'alerte mis en place et à l'utilisation de cette information par les pouvoirs publics.

Pour terminer, ce mémoire en vue de l'habilitation à diriger des recherches nous a permis de situer plus précisément nos objets d'étude en sciences de l'information et de la communication. Il nous a donné l'occasion de nous positionner sur les plans théoriques et méthologiques et d'identifier les perspectives de recherche sur l'organisation des connaissances.

Bibliographie

Agostinelli Serge (2013), « Connaissance : pseudo-concept partiellement opératoire », *Études de communication*, 2013 (39), p. 64-76.

Aguiar Fernando et Beigbeder Michel (2004), « Construction et utilisation de contextes autour des noeuds d'un hypertexte pour la recherche d'information », *Document numérique,* 2004, 8 (3), p. 71-82.

Amar Muriel (2004), « L'indexation aujourd'hui », *Les dossiers de l'ingénierie éducative : La fonction documentaire au cœur des TIC*, décembre 2004 (49), p. 61-65.

Amar Muriel (2000), *Les fondements théoriques de l'indexation. Une approche linguistique*. Paris : ADBS Editions, 2000, 355p.

Amos David (2005), *Organisation des connaissances dans les systèmes d'informations orientés utilisation. Contexte de veille et d'intelligence économique, Actes du colloque international de ISKO-France des 28-29 avril 2005*, Presses Universitaire de Nancy, 380p.

Antoniadis George (1984), *Elaboration d'un analyseur morpho-syntaxique d'une langue naturelle. Application à l'informatique documentaire*, Thèse de doctorat en informatique, sous la direction de Jacques Rouault, Université Grenoble 2, 1984.

Apothéloz Denis et Reichler-Béguelin Marie-Josée (1995), « Construction de la référence et stratégies de désignation » In Berrendonner Alain et Reichler-Béguelin Marie-Josée (ed.), 1995, p. 227-271.

Aronoff Mark (1976), *Word Formation in Generative Grammar*, Cambridge, Massachusets : The MIT Press, 1976, 134p. (Linguistic Inquiry Monographs one)

Arquembourg Jocelyne (2006), « De l'événement international à l'événement global : émergence et manifestations d'une sensibilité mondiale », In Arquembourg Joceylyne, Lochard Guy et Mercier Arnaud (coord.), *Evénements mondiaux regards nationaux : Hermès,* 2006 (46), CNRS Editions, p 13-21.

Auroux Sylvain (1998), *La raison, le langage et les normes*, Paris : PUF, 1998, 337p.

Aussenac-Gilles Nathalie et Condamines Anne (2007), « Corpus et terminologie », In Pédauque Roger T., *La redocumentarisation du monde*, Toulouse : Cépaduès Editions, 2007, p.131-147.

Bachimont Bruno (2007), *Ingénierie des connaissances et des contenus. Le numérique entre ontologies et documents.* Paris : Hermès Science Publications - Lavoisier, 2007, 279p. (Collection Science informatique et SHS)

Balicco Laurence, Bertier Marc, Clavier Viviane, Mounier Evelyne, Paganelli Céline et Staii Adrian (2007), *Les pratiques informationnelles des médecins du CHU de Grenoble*, Rapport de recherche, *Projet NOESIS. CHU de Grenoble*. Document non publié consultable à la Bibliothèque Yves-de-la-Haye, Université Stendhal.

Balicco Laurence (1993), *Génération de répliques en français dans une interface homme-machine en langue naturelle*. Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

Battaïa Céline (2013), *L'émotion dans les forums de discussion : structuration et évalaution de l'information de santé*. Thèse de doctorat en sciences de l'information et de la communication sous la direction conjointe de Laurence Balicco et Viviane Clavier, Université de Grenoble, 2013.

Bawden David (2006), « Users, user studies and human information behaviour. A three decade perspective on Tom Wilson's "On user studies and information needs" », *Journal of Documentation*, 2006, 62 (6), p. 671-679.

Beaulieu Micheline (2003), « Approaches to user-based studies in information seeking and retrieval : a Sheffiled perspective », *Journal or Information Science*, 2003, 29 (4), p. 239-248.

Beesley Kenneth and Karttunen Lauri (2003), *Finite State Morphology*, Stanford University: CSLI Publications, 2003, 509p. (Studies in computational linguistics).

Bellot Patrice, Cauvet Corine, Pasi Gabriella et Parlangeau Nathalie (dir.) (2012), *Approches pour la recherche d'information en contexte,* Paris : Hermès Science Publications - Lavoisier 2012, 146p.

Benzécri Jean-Paul (dir.) (1973), *L'analyse des données*, Paris, Bruxelles, Montréal : Dunod, 1973, 620p.

Berrendonner Alain (1995), « Redoublement actantiel et nominalisations », in *Problèmes de sémantique et de relations entre micro- et macrosyntaxe*, *Actes des rencontres de linguistique BENEFRI-Strasbourg, Neuchâtel,* 19-21 mai 1994, *Scolia*, n°5, Strasbourg : Pressses de l'Université de Strasbourg, 1995, p. 215-245.

Berrendonner Alain (1987), « La logique du soupçon », *Revue européenne des sciences sociales : Pensée Naturelle Logique et langage. A Jean-Blaise Grize,* 1987, XXV (77), Genève : Librairie Droz, p. 287-297.

Berrendonner Alain (1983a), *Cours critique de grammaire générative*, Lyon : Presses universitaires de Lyon, Fribourg : Editions, 1983, 320p. (Linguistique et sémiologie).

Berrendonner Alain (1983b), *Grammaire pour un analyseur. Aspects morphologiques*, *Cahiers du CRISS*, novembre 1990 (15), Grenoble.

Berrendonner Alain, Fredj Mounia, Oquendo Flavio et Rouault Jacques (1992), « Un système inférentiel orienté objet pour des applications en langue naturelle », in *Proceeding of the 14th conference on Computational linguistics*, *COLING'92*, 1992, n°2, p. 461-467.

Berrendonner Alain et Clavier Viviane (1997), « Examen d'une série morphologique dite "improductive" en français : les noms dénominaux en -age », *Actes des 1ères Rencontres du Forum de Morphologie, Mots possibles et mots existants*, 28-29. avril 1997, p. 35-45.

Berrendonner Alain et Miéville Denis (ed.) (1997), *Logique, discours et pensée : mélanges offerts à Jean-Blaize Grize*, Berlin, New-York, Paris : Peter Lang, 1997, 444p.

Berrendonner Alain et Reichler-Béguelin Marie-Josée (ed.) (1995), *Du syntagme nominal aux objets-de-discours : SN complexes, nominalisations, anaphores*, N° spécial des TRANEL 23, Neuchâtel : Université de Neuchâtel Institut de linguistique, 1995, 315p. (Travaux Neuchâtelois de Linguistique).

Besançon Romaric (2004), Ch. 2. « Technologies statistiques pour la recherche d'informations : les modèles vectoriels », In Ihadjadene M. (dir.), *Les systèmes de recherche d'informations : modèles conceptuels*, p. 35-54, Paris : Hermès Science Publications – Lavoisier, 2004.

Biber Douglas (1993), « Using register-diversified corpora for general language studies », *Computational Linguistics*, 1993, 19 (2), p.243-258.

Biber Douglas (1988), *Variation across Speech and Writing*, Cambridge, New-York : Cambridge University Press, 1988.

Bisseret André, Sebillote Suzanne et Falzon Pierre (1999), *Techniques pratiques pour l'étude des activités expertes.* Toulouse : ed. Octares, 1999, 155p. (Collection Travail)

Bonnafous Simone et Jost François (2000), « Analyse de discours, sémiologie et tournant communicationnel », *Réseaux,* 2000, 18 (100), p. 523-545.

Boubée Nicole et Tricot André (2010), *Qu'est ce que rechercher de l'information ? Etat de l'art*, Villeurbanne : Presses de l'Enssib, 2010, 287p. (Papiers, Série Usages des documents)

Boughanem Mohand et Savoy Jacques (2008), *Recherche d'information. Etat des lieux et perspectives*, Paris : Hermès Science Publications - Lavoisier 2008, 343p. (Hermès Science Publications).

Bouillon Pierrette (1998), *Traitement automatique des langues naturelles,* Paris, Bruxelles : Duculot, Paris : Aupelf-Uref, 1998, 245p.

Bourquin Jacques (1975), *La dérivation suffixale. Histoire d'une théorisation et de ses implications pédagogiques*. Thèse de doctorat en linguistique, Université de Besançon, 1975.

Bouvier-Ajam Laurent (2007), *Europeana. Etude sur les usages et les attentes relatifs à l'interface de consultation de la future Bibliothèque numérique Européenne*, Rapport final, 2007, 53 p. [En ligne] URL : http://www.bnf.fr/documents/ourouk.pdf Consultée le 5 mai 2013.

Braunwarth Michel, Mekaouche Abdelouahab et Bassano Jean-Claude (1994), « Information Retrieval System using Distributed Artificial Intelligence Tools », in *Intelligent Multimedia Information Retrieval Systems and Management, RIAO 94,* Rockefeller University, New-York, CID/CASIS, octobre 1994, p. 800-803.

Branca-Rosoff Sonia (1999), « Types, modes et genres : entre langue et discours » *Langage et société*, 1999 (87), p. 5-24.

Broudoux Evelyne (2013), « Quelles lectures du tagging ? Modélisation, techniques et usages », *Document numérique*, 2013, 16 (1), p. 55-71.

Broughton Vanda; Hansson Joacim; Hjørland Birger and López-Huertas, Maria J. (2005), Ch. 7 « Knowledge Organization », in Kajberg L. & Lørring L. (ed.), *European Curriculum Reflections on Library and Information Science Education*, p. 133-148, *Report of working group on LIS-education in Europe. Working seminar held in Copenhagen 11-12 August 2005 at the Royal School of Library and Information Science,* Copenhagen: Royal School of Library and Information Science. [En ligne]
URL : http://arizona.openrepository.com/arizona/bitstream/10150/105851/1/KnowledgeOrg_chapter%25207.pdf Consultée le 23 octobre 2013.

Brunet Etienne (2003), « Peut-on mesurer la distance entre deux textes ? », *Corpus*, 2003 (2). [En ligne] URL : http://corpus.revues.org/30. Consultée le 17 mai 2013.

Case Donald O. (2002), *Looking for Information: a Survey of Research on Information Seeking, Needs and Behaviour*, Bingley : Emerald, 3rd edition, 2002, 491p. (Library and Information Science)

Chanet Catherine (1996), *La demande dans le dialogue finalisé : de la surface linguistique aux représentations de l'action.* Thèse de doctorat en sciences de l'information et de la communication sous la direction de Jacques Rouault, Université Grenoble3, 1996.

Charaudeau Patrick et Maingueneau Dominique (dir.) (2002), *Dictionnaire d'analyse du discours,* Paris : Editions du Seuil, 2002, 661p.

Chartron Ghislaine, Dalbin Sylvie, Monteil Marie-Gaëlle, Verillon Monique (1989), « Indexation manuelle et indexation automatique. Dépasser les oppositions. », *Documentaliste*, juillet- octobre 1989, 26 (4-5), p. 181-187.

Chaudiron Stéphane (2004a), Ch. 8 « L'évaluation des systèmes de recherche d'informations », in Ihadjadene M. (dir.) *Les système de recherche d'informations : modèles conceptuels*, p. 185-207, Paris : Hermès Science Publications – Lavoisier, 2004.

Chaudiron Stéphane (2004b), Ch. 12 « La place de l'usager dans l'évaluation des systèmes de recherche d'informations », in Chaudiron Stéphane (dir.), *Evaluation des systèmes de traitement de l'information*, p 287-310, Paris : Hermès Science Publications - Lavoisier, 2004.

Chaudiron Stéphane et Ihadjadene Madjid (2010), « De la recherche de l'information aux pratiques informationnelles », *Études de communication*, 2010 (35), [En ligne] URL : http://edc.revues.org/index2257.html7 Consultée le 2 novembre 2012.

Chaumier Jacques (2003), *Les techniques documentaires au fil de l'histoire. 1950-2000*. Paris : ADBS Editions, 2003.

Chaumier Jacques (1982), *Analyse et langages documentaires. Le traitement linguistique de l'information documentaire*, Paris : Entreprise Moderne d'Edition, 1982, 186p.

Chevallet Jean-Pierre (2004), Ch. 5 « Modélisation logique pour la recherche d'informations », in Ihadjadene M. (dir.) *Les système de recherche d'informations : modèles conceptuels*, p. 105-138, Paris : Hermès Science Publications – Lavoisier, 2004.

Chiaramella Yves et Mulhem Philippe (2007), « La recherche d'information. De la documentation automatique à la recherche d'information en contexte », *Document numérique*, 2007, 10 (1), p. 11-38.

Chomsky Noam (1957), *Structures syntaxiques*, Paris : Ed. du Seuil, 144p.

Clavier Viviane et Paganelli Céline (dir.) (2013), *L'information professionnelle*, Paris : Lavoisier Hermès Sciences Publications, 243p. (Collection systèmes d'information et organisations documentaires)

Clavier Viviane (2013), Ch. 2 « L'information professionnelle dans les discours : le parent pauvre de l'information spécialisée », p. 47-69. In Clavier Viviane et Paganelli Céline (dir.) (2013), *L'information professionnelle*, Paris : Lavoisier Hermès Sciences Publications.

Clavier Viviane et Paganelli Céline (2012), Ch.10 « L'indexation de discours scientifiques : prise en compte des connaissances liées au positionnement de l'auteur », p. 171-181. In Mustapha El Hadi Widad (dir.) (2012), *L'organisation des connaissances : dynamisme et stabilité*, Cachan : Lavoisier - Hermès Sciences Publications (Traité des sciences et techniques de l'information).

Clavier Viviane et Paganelli Céline (2012), « Including Authorial Stance in the Indexing of Scientific Documents », *Knowledge Organization,* 39 (4), Würzburg : Ergon Verlag, p. 292-300.

Clavier Viviane et Paganelli Céline (2010), « De la consultation de documents scientifiques à leur indexation : pertinence de la notion de positionnement en sciences de l'information », *Les Enjeux de l'information et de la communication*, Supplément 2010B, 23p. [En ligne] http://w3.u-grenoble3.fr/les_enjeux/2010-supplementB/Clavier/index.html Consultée le 13 janvier 2013.

Clavier Viviane, Manes-Gallo Maria-Caterina, Mounier Evelyne, Paganelli Céline, Romeyer Hélène et Staii Adrian (2010), Ch. 19 «Dynamiques interactionnelles et rapports à l'information dans les forums de discussion médicale », p. 297-312. In Millerand Florence, Proulx Serge et Rueff Julien (dir.) (2010), *Le web social : mutation de la communication*, Presses Universitaires du Québec.

Clavier Viviane (2010), « Indexer des parcours thématiques pour valoriser les collections de presse numérisée », In Ihadjadene Madjid, Zacklad Manuel, Zreik Khaldoun (dir.) (2010), A*ctes du 13ème Colloque International sur le Document Electronique (CIDE.13)*, *Document numérique. Entre permanence et mutation*, 16-17 décembre 2010, INHA, Paris : Europia, p. 101-118.

Clavier Viviane et Romeyer Hélène (2008), « Travailler sur les discours en SIC : le cas des discours sur le cancer », *Actes du 16e congrès de la Société française des sciences de l'information et de la communication*, 11-13 juin 2008, 10p. [En ligne] http://www.sfsic.org Consultée le 13 janvier 2013.

Clavier Viviane (2006), « Le genre comme point d'accès au document : analyse comparée de textes scientifiques en mécanique et linguistique », *Journée de l'ATALA Typologies de textes pour le traitement automatique*, 9 décembre 2006, 5p. [En ligne] http://www.atala.org/article.php3?id_article=312 Consultée le 13 janvier 2013.

Clavier Viviane et Lafont-Terranova Jacqueline (2004), « Le français dans une filière technologique : une approche transversale pour l'apprentissage des discours de spécialité », *Actes du 9ème colloque international de l'AIRDF, Le français : discipline singulière, plurielle ou transversale?* Québec, 26-27-28 août 2004, publication sur CD-ROM, 14p.

Clavier Viviane et Guet Jean-Michel (2004), « Produire des écrits scientifiques et techniques : de la construction des savoirs en sciences des matériaux à leur appropriation. » *Actes du Colloque National de la Recherche dans les IUT*, 6-7 mai 2004, tome 2, p. 95-102.

Clavier Viviane, Cleuziou Guillaume et Martin Lionel (2002), « Organisation conceptuelle de mots pour la Recherche d'Information sur le Web », *Actes de la Conférence d'Apprentissage (CAP) 2002*, 17-19 juin 2002 à Orléans, Presses Universitaires de Grenoble, p. 220-235.

Clavier Viviane et Poudat Céline (2001), « Teaching Machine translation in non computer science subjects : report of an educational experience within the University of Orléans », *Actes du VIIIème congrès international de Traduction Automatique (Machine Translation Summit),* du 22 au 28 septembre 2001, Saint-Jacques de Compostelle, Espagne, p. 19-23.

Clavier Viviane (1999), « Morphologie et structuration des connaissances : pour une intégration d'outils linguistiques dans la recherche d'information en texte intégral ». In Maniez Jacques, Mustafa El Hadi Widad (1999) (dir), *Actes des Premières Journées du Chapitre Français de l'ISKO, Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et recherche d'information*, Edition du Conseil Scientifique de l'Université Charles De Gaulle Lille 3, p. 385-386. (Collection UL3: Travaux & Recherches)

Clavier Viviane et Coret Muriel (1998), « Dérivation avec et sans suffixe ». In Caron Bernard (Ed.) (1998), *Proceedings of the 16th International Congress of Linguists (Paris, July 20-25, 1997)*, publication sur CD-ROM, Oxford : Pergamon Press, 5000p.

Clavier Viviane (1998), *Etude sémantique des noms dérivés de verbe : problèmes d'aspect*, Rapport final du stage post-doctoral destiné à l'AUPELF-UREF, Université de Fribourg.

Clavier Viviane, Froissart Christel et Paganelli Céline (1997), « Objects and actions : two concepts of major interest for information retrieval in full-text data basis », *Proceedings NLDB'97, Third Workshop on Applications of Natural Language to Information System,* Vancouver, June 25-27 1997.

Clavier Viviane et Coret Muriel (1997), « Un dictionnaire électronique pour la reconnaissance des formes dérivées », *Meta, Journal des traducteurs*, Numéro spécial double *Lexicologie et Terminologie* sous la direction d'André Clas, 42 (2), Les Presses de l'Université de Montréal, p. 307-316.

Clavier Viviane (1997) « Retrouver l'information dans de gros corpus techniques : contribution de la morphologie à l'extraction de connaissances », *Affiche aux Rencontres de Linguistique Appliquée organisées par l'AFLA*, *Construction et utilisation de grands corpus*, Université Paris 7, 24-27 septembre 1997.

Clavier Viviane, (1996), *Modélisation de la suffixation pour le traitement automatique du français. Application à la recherche d'information*, Thèse de Doctorat en Sciences de l'Information et de la Communication, 1996, Université Stendhal, Grenoble.

Clavier Viviane (1996), « Morphologie et reconnaissance des mots inconnus en TAL : compte-rendu de l'évaluation Grâce », *Actes GDR-PRC Séminaire Lexique, Représentations et Outils pour les Bases Lexicales, Morphologie Robuste*, 13-14 novembre 1996, p. 97-106.

Clavier Viviane, Warren Karine, Lallich-Boidin Geneviève et Stéfanini Marie-Hélène (1996), « Intégration de la morphologie dérivationnelle dans un système distribué d'analyse du français écrit », *Actes du colloque Informatique & Langue Naturelle (ILN'96),* 9-10 octobre 1996, Université de Nantes, pp. 103-120.

Clavier Viviane, Lallich-Boidin Geneviève, Rouault Jacques et Timimi Isamaïl (1995), « Analyse automatique de discours : perspectives 1995 », in Bolasco Sergio, Lebart Ludovic, Salem André (Ed.) (1995), *Actes JADT'1995*, *3rd International Analysis of Textual Data*, 11-13 décembre 1995, Rome, p. 163-172.

Clavier Viviane et Lallich-Boidin Geneviève (1994), « Modélisation linguistique de la suffixation en vue de l'analyse automatique », *TAL,* 35 (2), p. 129-143.

Clavier Viviane (1990), Dictionnaire syntactico-sémantique des verbes du corpus MMI2. Interface multi-mode pour une interaction homme-machine avec une base de connaissances, Rapport d'activité du contrat ESPRIT-MMI2, CRISS, octobre 1990.

Clavier Viviane (1990), *Morphologie dérivationnelle des substantifs déverbaux*, Mémoire de DEA en sciences du langage, sous la direction de Geneviève Lallich-Boidin, Université Stendhal, Grenoble 3, 1990.

Clavier Viviane (1989), Untersuchung von Ableitungsregeln der bar-Adjektive im Hinblick auf maschinelle Übersetzung, Mémoire de Maîtrise (trad.) Etude des règles dérivationnelles des adjectifs en -bar en vue de la traduction automatique, mémoire de maîtrise d'allemand sous la direction conjointe de Roger Sauter et Geneviève Lallich-Boidin, Université Jean Monnet, 1989, Saint-Etienne.

Cleuziou Guillaume, Martin Lionel, Clavier Viviane et Vrain Christel (2004a), «DDOC: Overlapping Clustering of Words for Document Classification», In Apostolico, Alberto; Melucci, Massimo (Eds.), (2004) *Proceedings, Lecture Notes* in *Computer Science*, *String Processing and Information Retrieval*, Springer, Vol. 3246, 2004, XIV, 11th International Conference, SPIRE 2004, Padova, Italy, October 5, p. 127-128.

Cleuziou Guillaume, Clavier Viviane, Martin Lionel, Vrain Christel (2004b), « Regroupement d'attributs en classes non-disjointes. Quel impact sur la classification de documents ? », *Workshop Fouille de textes, 4ème journée d'Extraction et Gestion des Connaissances (EGC'2004)*, Clermont Ferrand, janvier 2004. [En ligne] http://www.univ-orleans.fr/lifo/Members/cleuziou/papers/CMCV_EGCFT_04.pdf Consultée le 13 janvier 2013.

Cleuziou Guillaume (2004), *Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information*, Thèse de doctorat en informatique sous la direction de Christel Vrain, Université d'Orléans, 2004.

Cleuziou Guillaume, Clavier Viviane et Martin Lionel (2003), « Structuration d'unités textuelles. Une méthode de regroupement fondée sur la recherche de cliques dans un graphe de cooccurrences » *Actes de la Conférence Terminologie et Intelligence Artificielle* (TIA'2003), Juin 2003, p 179-182. [En ligne] http://tia.loria.fr/TIA/IMG/pdf/TIA2003.pdf Consultée le 13 janvier 2013.

Cleverdon, Cyril W. (1967), « The Cranfield Tests on Index Language Devices », in *Aslib Proceedings*, 19 (6), p. 173-194.

Condamines Anne et Poibeau Thierry (2008), « Présentation », *Revue Française de Linguistique Appliquée,* 2008, XIII (1), p. 5-8.

Corbin Danielle (1991), « La formation des mots : structures et interprétations », *Lexique,* 1991 (10), Lille : Presses Universitaires de Lille, p. 7-30.

Corbin Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, 2 vol. *Linguistische Arbeiten*, 1987 (193), Tübingen : Max Niemeyer Verlag.

Corblin Francis (1995), *Les formes de reprise dans le discours : anaphores et chaînes de références,* Rennes : Presses universitaires de Rennes, 1995.

Coret Muriel (1994), *Problèmes de suffixation et structuration du lexique. Etude des mots en –eur, -age, -ment, -ion.* Thèse de doctorat en linguistique sous la direction d'Hélène Huot, Université Paris 7, Denis Diderot, 1994.

Cori Marcel (2008), « Des méthodes de traitement automatique aux linguistiques fondées sur les corpus », *Langages*, 2008, 1(171), p. 95-110.

Cori Marcel et Léon Jacqueline (2002), « La constitution du TAL. Étude historique des dénominations et des concepts », *TAL*, 2002, 43 (3), p. 21-55.

Courbières Caroline (2002), « Une approche communicationnelle de l'analyse documentaire », in Couzinet V. et Régimbeau G. (dir.), *Recherches récentes en sciences de l'information : convergences et dynamiques*, *Actes du colloque international LERASS-MICS*, 21-22 mars 2002, Toulouse, Université Paul Sabatier, Paris : ADBS, p. 105-125.

Courtine Jean-Jacques (1981), « Analyse du discours politique », *Langages,* juin 1981 (62), p. 9-128.

Courtois Blandine (1994-95), « Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL », *Cahiers du CIEL 1994-1995* : *Théories et pratiques du lexique*, Université Paris 7, p. 87-107. [En ligne] URL : http://infolingu.univ-mlv.fr/english/Bibliographie/Articles/3Courtois.pdf. Consultée le 13 janvier 2013.

Courtois Blandine et Silberztein Max (1989), « Les dictionnaires électroniques DELAS et DELAC », In *RELAI: Recherches en Linguistique Appliquée à l'Informatique. Actes du colloque "La description des langues naturelles en vue d'applications informatiques",* Québec : Université Laval, 1988.

Coutaz Joëlle, Calvary Gaëlle, Demeure Alexandre et Balme Lionel (2012), Ch. 9 « Systèmes interactifs et adaptation centrée utilisateur : la plasticité des Interfaces Hommes-Machine ». In Calvary G., Delot T., Sèdes F. et Tigli J.-Y. (2012), *Informatique et Intelligence ambiante : des capteurs aux applications*, 57 p. Paris : Hermès Science Publications - Lavoisier, 2012. (Traité IC2, série Informatique et Systèmes d'Information) [En ligne] URL : http://iihm.imag.fr/publs/2012/LivreAmi-Chap9-Plasticite-CoutazCalvary .pdf Consultée le 7 mars 2013.

Couzinet Viviane (2012), Ch. 1 « L'organisation des connaissances en regard des sciences de l'information et de la communication, une exception française », in Mustapha El Hadi W. (dir.), *L'organisation des connaissances : dynamisme et*

stabilité, p. 35-50, Cachan : Hermès Sciences Publications - Lavoisier, 2012 (Traité des sciences et techniques de l'information).

Couzinet Viviane (2011), « Questions des dispositifs info-communicationnels », in Gardiès C. (dir.), *Approche de l'information documentation : concepts fondateurs*, p.117-130, Toulouse : Éditions Cépadues, 2011.

Couzinet Viviane (dir.) (2009), *Dispositifs info-communicationnel : questions de médiations documentaires,* Paris : Hermès Sciences Publications – Lavoisier, 2009, 263 p (Systèmes d'information et organisations documentaires)

Couzinet Viviane (2006), « Les connaissances au regard des sciences de l'information et de la communication : sens et sujets dans l'inter-discipline » *Semaine de la connaissance*, vol. 1, Université de Nantes, 26-30 juin 2006, 6p. [En ligne]www.irit.fr/SDC2006/cdrom/contributions/Couzinet_SDC2006.pdf Consultée le 12 avril 2013.

Couzinet Viviane (2005), « Intelligence économique et sciences de l'information et de la communication : quelles questions de recherche ? », in Amos D. (coord.), *Organisation des connaissances dans les systèmes d'informations orientés utilisation. Contexte de veille et d'intelligence économique*, *Actes du colloque international de ISKO-France*, p. 13-25, Presses Universitaire de Nancy, 2005.

Couzinet Viviane (2002), « Convergences et dynamiques nationales : pour une mise en visibilité des recherches en sciences de l'information », in Couzinet V., Régimbeau G. (dir.), *Recherches récentes en sciences de l'information. Convergences et dynamiques*, *Actes du colloque international LERASS-MICS*, 21-22 mars 2002, Toulouse, Université Paul Sabatier, p. 9-14, Paris : ADBS, 2002.

Cox Andrew M. (2012), « An exploration of the practice approach and its place in information science », *Journal of Information Science*, April 2012, 38 (2), p. 176-188.

Cox Andrew M. (2008), « An exploration of concepts of community through a case study of UK university web production », *Journal of Information Science*, 2008, 34 (3), p. 327-345.

Daille Béatrice et Romary Laurent (dir.) (2001), *Traitement automatique des langues : Linguistique de corpus,* 1991, 2 (42).

Dal Georgette et Namer Fiammetta (2000), « Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'information », *TAL*, 2000, 41 (2), p. 423-446.

Dalbin Sylvie et Guyot Brigitte (2007), « Documents en action dans une organisation : des négociations à plusieurs niveaux », *Études de communication*, 2007, 30, p. 55-70.

Dell François (1973), *Les règles et les sons : introduction à la phonologie générative.* Paris : Hermann, 1973, 286 p.

Dendale Patrick et Tasmowski Liliane (1994), « Présentation. L'évidentialité ou le marquage des sources du savoir. » In Dendale Patrick Tasmowski Liliane, *Langue Française : Les sources du savoir et leurs marques linguistiques*, 1994,102 (1), p. 3-7.

Dervin Brenda, (1983), *Information as a user construct : the relevance of perceived information needs to synthesis and interpretation* (pp. 153-183), in Ward S. A. et Reed L. J. (ed.), *Knowledge structure and use : implications for synthesis and interpretation*, p. 153-183, Philadelphia : Temple University Press, 1983.

Desprès-Lonnet Marie (2009), « L'écriture numérique du patrimoine, de l'inventaire à l'exposition : les parcours de la base Joconde », *Culture & Musées,* 2009 (14), p. 19-38.

Ducrot Oswald et Todorov Tzvetan (1972), *Dictionnaire encyclopédique des sciences du langage,* Paris : Edition du Seuil, 470p.

Durieux Valérie (2010), « Collaborative tagging et folksonomies », *Les Cahiers du numérique*, 2010, 6 (1), p. 69-80. [En ligne]
URL : www.cairn.info/revue-les-cahiers-du-numerique-2010-1-page-69.htm
Consultée le 16 juin 2013.

Eymard Gilbert (1992), *Traitement documentaire des sommaires : des mots-clés à l'extraction de connaissances. Application à une documentation technique.* Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Grenoble2, 1992.

Eymard Gilbert (1988), *L'interrogation en langue naturelle d'une base de données textuelle : un chantier. Application aux « Spécifications Technique d'Utilisation du Minitel 1B (STUM1B) »*, Mémoire de DEA, CRISS, Université de Grenoble2, 1988.

Finkelstein Lev, Gabrilovich Evgeniy, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, Ruppin Eytan (2002), « Placing Search in Context: The Concept Revisited », *ACM Transactions on Information Systems,* January 2002, 20 (1), p. 116–131.

Fluhr Christian (1992), « Le traitement du langage naturel dans la recherche d'information documentaire », *Interfaces Intelligences dans l'information scientifique et technique, cours INRIA*, Klingenthal (Bas-Rhin), mai 1992.

Fluhr Christian (1985), « Le traitement et l'interrogation des bases de données textuelles », Editions de l'université de Bruxelles, Editions Bruylant, 1985, p. 97-114.

Fluhr Christian (1977), *Algorithmes à apprentissage et traitement automatique des langues*, Thèse de doctorat en informatique, Orsay, Université de Paris Sud, 1977.

Fondin Hubert (2002), « La "science de l'information" et la documentation ou les relations entre science et technique », *Documentaliste-Sciences de l'Information*, 2002, 39 (3), p.122-129.

Fondin Hubert (2001), « La science de l'information : posture épistémologique et spécificité disciplinaire », *Documentaliste-Science de l'Information*, 2001, 38 (2), p.112-122.

Fradin Bernard (1993), *Organisation de l'information lexicale et interface morphologie / syntaxe dans le domaine verbal*. Thèse de doctorat d'état en linguistique sous la direction de Nicolas Ruwet, Université Paris 8, 1993.

Francis Elie et Quesnel Odile (2007), « Indexation collaborative et folksonomies », *Documentaliste-Sciences de l'Information*, 2007, 44 (1), p. 58-63.

Francony Jean-Marc (1993), *Modélisation du dialogue et représentation du contexte d'interaction dans une interface de dialogue multi-modes dont l'un des modes est dédié à la langue naturelle écrite*. Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

Fredj Mounia (1993), *Saphir. Un système d'objets inférentiels : contribution à l'étude des raisonnements en langue naturelle.* Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1993.

Froissart Christel (1992), *Robustesse des interfaces homme-machine en langue naturelle*. Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Pierre Mendès-France, Grenoble2, 1992.

Froissart Christel et Lallich-Boidin Geneviève (1996), « Morphologie robuste et analyse automatique de la langue : étude réalisée à partir des corpus de l'évaluation GRACE », in *Représentations et Outils pour les Bases Lexicales, Morphologie Robuste, Actes GDR-PRC Séminaire Lexique,* p. 13-14, novembre 1996, p. 88-96.

Fuchs Catherine (1994), *Paraphrase et énonciation,* Paris : Ophrys, 185p. (Collection L'Homme dans la Langue)

Fuchs Catherine (1993), *Linguistique et Traitements Automatiques des Langues*, Paris : Hachette, 1993, 303p. (Hachette université. Langue, linguistique communication)

Gardiès Cécile (dir.) (2011), *Approche de l'information-documentation. Concepts fondateurs*. Toulouse : Cépaduès Editions, 2011, 232p.

Gardiès Cécile et Fabre Isabelle (2012), « Définition et enjeux de la médiation numérique documentaire ». In Galaup Xavier (dir.), *Développer la médiation documentaire numérique,* p. 45-58, Villeurbanne : presses de l'Enssib, 2012.

Garfield Eugene (1997), « A tribute to Calvin N. Mooers, a pioneer of information retrieval », The Scientist, 1997, 11 (6), p. 9 [En ligne]
URL : http://www.answers.com/topic/calvinmooers Consultée le 12 février 2013.

Gaussier Eric et Stéfanini Marie-Hélène (dir.) (2003), *Assistance intelligente à la recherche d'informations*, Paris : Hermès Science Publications – Lavoisier, 2003, 319p. (Traité des sciences et techniques de l'information)

Gnoli Claudio (2012), Ch. 2 « Des métadonnées représentant quoi ? Etablissement d'une distinction entre les dimensions ontiques, épistémologiques et documentaires dans l'organisation des connaissances », in Mustapha El Hadi W. (dir.), *L'organisation des connaissances : dynamisme et stabilité*, p. 51-66., Cachan : Hermès Sciences Publications - Lavoisier, 2012 (Traité des sciences et techniques de l'information).

Grésillon Almuth (2006), *La critique génétique, aujourd'hui et demain*. Item [En ligne] URL : http://www.item.ens.fr/index.php?id=14174 Consultée le 13 mai 2013.

Grivel Luc (2011), *La recherche d'information en contexte : outils et usages applicatifs*, Paris : Hermès Sciences Publications - Lavoisier, 2011, 279 p. (Traité des sciences et techniques de l'information)

Gross Maurice (1975), *Méthodes en syntaxe. Régime des constructions complétives*, Paris : Hermann, 1975, 414p.

Gruaz Claude (1988), *La dérivation suffixale en français contemporain*, Mont-Saint-Aignan : Publication de l'Université de Rouen, 436p.

Guiraud Pierre (1954), Les *caractères statistiques du vocabulaire. Essai de méthodologie*, Paris : PUF, 1954, 116p.

Guyot Brigitte (2006), *Dynamiques informationnelles dans les organisations,* Paris : Hermès - Lavoisier, 2006, 236p. (Collection, Finance, Gestion, management)

Habert Benoit (2005), « Portrait de linguiste(s) à l'instrument » *Texto !,* 2005, X (4). [En ligne]URL : http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html Consultée le 23 juillet 2013.

Habert, Benoît (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? » In Bilger, M. (ed.), *Linguistique sur corpus. Etudes et réflexions*, Perpignan : Presses Universitaires de Perpignan. [En ligne]
URL : http://atala.biomath.jussieu.fr/je/010428/Habert/Perpignan00/node3.html Consultée en mai 2013.

Habert Benoit, Fabre Cécile, Issac Fabrice (1998), *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques,* Paris : InterEditions, 1998, 320p.

Habert Benoît, Illouz Gabriel, Lafon Pierre, Fleury Serge, Folch Helka, Heiden Serge et Prévost Sophie (2000), « Profilage de textes : cadre de travail et expérience », in *Ve Journées internationales d'analyse statistique des données textuelles* (JADT'2000), Lausanne. [En ligne]
URL : http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/56/56.pdf
Consultée le 25 mars 2013.

Habert Benoit, Nazarenko Adeline et Salem André (1997), *Les linguistiques de corpus,* Paris : Armand Colin, 1997, 240p.

Halle Morris (1973), « Prolegomena to a theory of Word Formation », *Linguistic Inquiry*, 1973, 4 (1), p. 3-16.

Harris Zellig (1963), *Discours Analysis Reprints*, The Hague : Mouton, 1963, 73p.

Helsloot Niels et Hak Tony (2000), « La contribution de Michel Pêcheux à l'analyse de discours », *Langage et société,* 2000*,* 1(91), p. 5-33.

Hjørland Birger (2008), « What is Knowledge Organization (KO)? », *Knowledge*

Ho-Dac Lydia-Mai, Lemarié Julie, Péry-Woodley Marie-Paule et Vergez-Couret Marianne (2012), « Multidisciplinary Perspectives on Signalling Text Organisation: Introduction to the Special Issue », *Discours*, 2012 (10). [En ligne] URL : http://discours.revues.org/8598. Consultée le 20 août 2013.

Hubert Gilles (2010), *Recherche d'information et contexte,* Habilitation à diriger des recherches en informatique, Université Paul Sabatier, Toulouse 3, 2010. [En ligne] URL : ftp://ftp.irit.fr/IRIT/SIG/2010_HDR_H.pdf Consultée le 12 février 2013.

Hudon Michèle et Mustafa el Hadi Widad (2013), « Introduction », *Études de communication* : *Organisation des connaissances : épistémologie, approches théoriques et méthodologiques*, 2012 (39), p. 9-14.

Hudon Michèle et Mustafa el Hadi Widad (2010), « Organisation des connaissances et des ressources documentaires », *Les Cahiers du numérique*, 2010, 6 (3), p. 9-38.

Ibekwe-SanJuan Fidelia (2012), « The French Conception of Information. Une exception française? », *Journal of the American Society for Information Science and Technology,* 2012, 63 (9), p. 1963-1709.

Ibekwe-SanJuan Fidelia (2007), *Fouille de textes : méthodes, outils et applications*, Paris : Hermès Science Publications – Lavoisier, 2007, 352p. (Collection Systèmes d'information et organisations documentaires)

Ibekwe-SanJuan Fidelia (2006), « Connaissances terminologiques et systèmes d'informations stratégiques », in Mustafa El Hadi W. (dir.), *Terminologie et accès à l'information*, p. 187-210, Paris : Hermès Science Publications – Lavoisier, 2006. (Traité des Sciences et techniques de l'information)

Ihadjadene Madjid (dir.) (2004), *Les systèmes de recherche d'information. Modèles conceptuels*. Paris : Hermès Science Publications - Lavoisier, 2004, 216p. (Traité des sciences et techniques de l'information)

Ihadjadene Madjid et Chaudiron Stéphane (2010), « Quels modèles pour analyser l'accès à l'information dans les organisations ? », *Les Enjeux de l'Information et de la Communication*, Supplément 2010B. [En ligne]

URL : http://w3.u-grenoble3.fr/les_enjeux/2010-supplementB/ChaudironIhadjadene/index.html. Consultée le 2 février 2013

Ihadjadene Madjid et Chaudiron Stéphane (2008), Ch. 7 « L'étude des dispositifs d'accès à l'information électronique : approches croisées » in Papy F. (dir.), *Problématiques émergentes dans les sciences de l'information*, p. 183-207, Paris : Hermès Science Publications – Lavoisier, 2008.

Ihadjadene Madjid et Favier Laurence (2008), « Langages documentaires : vers une crise d'autorité » in Couzinet V., Chaudiron S. (coord.), *L'organisation des connaissances à l'ère numérique, Sciences de la société*, 2008 (75), p. 11-22.

Ingwersen Peter and Järvelin Kalervo (2005), *The turn. Integration of Information Seeking and Retrieval in Context*, Dordrecht : Springer, 2005, 448p.

Ingwersen Peter (1992), *Information Retrieval Interaction*, London : Taylor Graham, 1992, 246p.

Jacquemin Christian (dir.) (2000), « Présentation », *TAL* : *Traitement automatique des langues pour la recherche d'information*, 2000, 41 (2), p. 327-331.

Jaquemin Christian et Zweigenbaum Pierre (2000), Ch. 4 « Traitement automatique des langues pour l'accès au contenu des documents », In Le Maître J., Charlet J.et Garbay C. (éd.), *Le document multimedia en sciences du traitement de l'information*, p. 71–109, Toulouse : Cépaduès, 2000.

Jeanneret Yves (2000), *Y a-t-il (vraiment) des technologies de l'information ?,* Villeneuve d'Ascq :Presses du Septentrion, 2000, 134p.

Jing Hongyan et Tzoukermann Evelyne (1999), « Information Retrieval Based on context Distance and Morphology », *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'99*, New-York : ACM, p. 90-96.

Kessler Brett, Nunberg Geoffrey and Schültze Hinrich (1997), « Automatic detection of text genre », Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'97), 1997, p. 32-38.

Koskenniemi Kimmo (1983), *Two-Level Morphology : A general Computational Model for Word-Form Recognition and Production,* Ph. D. thesis, University of Helsinki, 1983.

Kovacs Susan et Timimi Ismaïl (2006), « Bonnes feuilles. Indice, index, indexation », *Documentaliste-Sciences de l'Information*, 2006, 43 (3), p. 210-215.

Lainé-Cruzel Sylvie (1999), « PROFILDOC. Filtrer une information exploitable », *BBF*, 1999, 44 (5), p. 60-64. [En ligne]
URL : http://bbf.enssib.fr/consulter/bbf-1999-05-0060-010 Consultée le 02 mai 2013.

Lallich-Boidin Geneviève (1986), *Analyse syntaxique automatique du français écrit : applications à l'indexation automatique*, thèse de doctorat soutenue en informatique, sous la direction de Jacques Rouault, Université Grenoble2, 1986.

Lallich-Boidin Geneviève et Maret Dominique (2005), *Recherche d'information et traitement de la langue : fondements linguistiques et applications,* Villeurbanne : Presses de l'Enssib, Novembre 2005, 288p. (Les Cahiers de l'Enssib)

Landron, Pierre-Yves (2010), « Valoriser la presse illustrée du XIXème : l'exemple de la BM Lyon », Communication présentée aux Journées d'études *Regards croisés sur la mise en ligne et la valorisation de la presse XIX-XXI*, Cluster 13 « Culture, patrimoine et création » les 6 et 7 mai 2010 à la Bibliothèque Municipale de Lyon.

Lebrave Jean-Louis (2007), « *Les manuscrits entre la substance de l'expression et la substance du contenu*. L'Item, 2007. [En ligne]
URL : http://www.item.ens.fr/index.php?id=187198 Consultée le 13 mai 2013.

Lee Yong-Bea and Myaeng Sung Hyon (2002), « Text genre classification with genre-revealing and subject-revealing features », in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'02, ACM Press, 2002, p. 145-150.

Le Guern Michel (1983), « Sémantique et syntaxe des descripteurs », *Communication à l'Ecole d'été des Sciences de l'information*, Vignieu, sept. 83, DBMIST, 1983.

Le Marec Joëlle (2004), Ch. 15 « Les études d'usage et leur prise en compte dans le champ culturel », in Chaudiron S. (dir.), *Evaluation des systèmes de traitement de l'information*, p. 353-372, Paris : Hermès Science Publications – Lavoisier, 2004.

Le Marec Joëlle (1997), Article « Sociologie des pratiques informationnelles », in Cacaly S., *Dictionnaire encyclopédique de l'information et de la documentation*, sous la direction. Paris : Nathan, 1997, 634p.

Léon Jacqueline (2010), « AAD69 : archéologie d'une étrange machine », *Semen,* 2010 (29). [En ligne] URL : http://semen.revues.org/8823 Consultée le 29 juillet 2013.

Léon J.acqueline (2001), « Conceptions du mot et débuts de la traduction automatique », *Histoire Epistémologie Langage*, 2001, 23 (1), p. 81-106.

Leriche Françoise et Meynard Cécile (2008), « Introduction. De l'hypertexte au manuscrit : le manuscrit réapproprié. Enjeux, expérimentations, perspectives. », in Leriche F. et Meynard C. (coord.), *De l'hypertexte au manuscrit. L'apport et les limites du numérique pour l'édition et la valorisation de manuscrits littéraires modernes*, p. 9-36, Grenoble : Ellug, 2008, 305p.

Lerat Pierre (2005), « Terme et microcontexte. Les prédications spécialisées » *Actes du colloque* « Mots, termes et contextes » (Bruxelles, 8-10 septembre 2005). [En ligne]

URL : http://perso.univ-lyon2.fr/~thoiron/JS%20LTT%202005/pdf/Lerat.pdf
Consultée le 2 novembre 2012.

Lesquins Noémie (2007), *Europeana : rapport de bilan sur les usages et les attentes des utilisateurs*, Bibliothèque Nationale de France, 2007, 60 p. [En ligne]
URL : http://www.bnf.fr/documents/europeana_2007.pdf Consultée le 6 janvier 2013.

Liquète Vincent (2011), *Des pratiques d'information à la construction de connaissances en contexte : de l'analyse à la modélisation SEPICRI (Systèmes, Environnement, Pratiques Individuelles, Collectives et Représentations de l'Information),* Mémoire d'habilitation à diriger des recherches, Université de Rouen, 2011.

Liquète Vincent, Fabre Isabelle et Gardiès Cécile (2010), « Faut-il reconsidérer la médiation documentaire ? », *Les Enjeux de l'Information et de la Communication*, Supplément 2010B. [En ligne] URL : http://w3.u-grenoble3.fr/les_enjeux/2010-dossier/Liquete-Fabre-Gardies/Liquete-Fabre-Gardies.pdf Consultée le 16 avril 2013.

Lovins Julie B. (1968), « Development of a stemming algorithm », *Mechanical Translation and Computational Linguistics,* 1968, 11 (1), p 22-31.

Maingueneau Dominique (2008), « Analyse du discours et littérature : problèmes épistémologiques et institutionnels », *Argumentation et Analyse du Discours*, 2008 (1). [En ligne] URL : http://aad.revues.org/351 Consultée le 19 juillet 2013.

Maingueneau Dominique (1991), *L'analyse de discours. Introduction aux lectures de l'archive*, Paris : Hachette Supérieur, 1991, 268p. (Hachette Université Linguistique)

Maniez Jacques (1994), *Actualité des langages documentaires : fondements théoriques de la recherche d'information*, Paris : ADBS, 1994, rééd. en 2002, 395p. (Série Etudes et techniques).

Matharan Judith, Chaguiboff Jean et Alliot François (2008), *Rapport d'étude sur les usages communautaires et collaboratifs, sur place et à distance, des ressources numérisées de la BnF,* Bibliothèque Nationale de France, 2008. [En ligne]
URL : http://www.bnf.fr/documents/rapport_web_communaute.pdf
Consultée le 13 mars 2013.

Maurel Dominique (2010), « *Sense-making* : un modèle de construction de la réalité et d'appréhension de l'information par les individus et les groupes », *Études de communication*, 2010 (35). [En ligne] URL : http://edc.revues.org/2306 Consultée le 30 janvier 2013.

McKenzie Pamela J. (2003), « A model of information practices in accounts of everyday life information seeking », *Journal of Documentation*, 2003, 59 (1), p. 19-40.

Macmurray Erin (2012), *Discours de presse et veille stratégique d'événements Approche textométrique et extraction d'informations pour la fouille de textes,* Thèse de doctorat en sciences du langage sous la direction d'André Salem, Université de la Sorbonne nouvelle, Paris III, 2012. [En ligne]URL : http://tel.archives-ouvertes.fr/tel-00740601 Consultée le 2 mai 2013.

Memmi Daniel (2000), « Le modèle vectoriel pour le traitement de documents », *Cahiers Leibniz,* 2000 (14), 30p. [En ligne]
URL : http://www.ieml.org/IMG/pdf/vectoriel.pdf Consultée le 12 février 2013.

Menon Bruno (2013), Ch. 5 « Quelles indexation pour l'information professionnelle ? », in Clavier V. et Paganelli C. (dir.), *L'information professionnelle*, p. 83-105, Paris : Hermès Sciences Publications - Lavoisier, 2013.

Metzger Jean-Paul (1988), *Syntagmes nominaux et information textuelle. Reconnaissance automatique et représentation*, Thèse de doctorat d'état es-sciences, sous la direction de Richard Bouché et Michel Le Guern, Université Claude Bernard, Lyon I, 1988.

Malrieu Denise et Rastier François (2001), « Genres et variations morphosyntaxiques », *Traitement Automatique des Langues*, 2001, 42 (2), p. 547-577.

Martin Robert (1995), « L'informatique et les dictionnaires de l'InaLF », in Pruvost J. (coord.), *Les Dictionnaires de langue française et l'informatique. Actes du colloque "La journée des dictionnaires"*, p. 31-39, Cergy-Pontoise : Centre de recherche Texte/Histoire, 1995.

Matthews Peter H. (1974), *Morphology. An introduction to the theory of word-structure,* Cambridge, London New-York : Cambridge University Press, 1974, 243p.

Merlo Aurélie (2012), « Système de prédiction de néologismes formels: le cas des N suffixés par –IER dénotant des artefacts », *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, vol. 3, RECITAL, p. 57–70, Grenoble du 4 au 8 juin 2012, ATALA & AFCP. [En ligne] URL : http://aclweb.org/anthology-new/F/F12/F12-3005.pdf Consultée le 12 juillet 2013.

Michel Christine et Lainé-Cruzel Sylvie (1999), « Profil-doc : un prototype de système de recherche d'information personnalisé selon le profil des utilisateurs. » In *Ateliers A2, 11ème conférence francophone Interaction Homme Machine IHM'99 :L'interaction pour tous*, Montpellier.

Michel Jean (2001), « Le knowledge management, entre effet de mode et (ré)invention de la roue... », *Documentaliste-Sciences de l'Information*, 2001, 38 (1), p. 176-186.

Miège Bernard (2007), *La société conquise par la communication*. *Les Tic entre innovation technique et ancrage social*, tome 3, Grenoble : PUG, 2007, 235p. (Communication, médias et sociétés)

Miège Bernard (2004), *L'information-communication, objet de connaissance*, Bruxelles : De Boeck, Paris : INA, 2004, 248p. (Collection Médias Recherche).

Miège Bernard (dir.) (2003), *Communication personnes systèmes informationnels*, Paris : Hermès Science Publications – Lavoisier, 2003, 196p. (Traité des sciences et techniques de l'information)

Montgomery Christine A (1972), « Linguistics and Information Science », *Journal of the American Society for Information Science and Technology,* May-Jun 72, 23 (3), p. 195-219.

Moreau Fabienne et Claveau Vincent (2006), « Extension de requêtes par relations morphologiques acquises automatiquement », *Actes de la 3ème Conférence en Recherche d'Informations et Applications, (CORIA '06),* p. 181-192, Mars 2006, Lyon, France.

Moreau Fabienne, Sébillot Pascale (2005), *Contributions des techniques du traitement automatique des langues à la recherche d'information,* Rapport de recherche n° 5484, INRIA, Rennes, Février 2005, 34p. [En ligne] URL : http://hal.inria.fr/inria-00070523 Consultée le 13 juin 2013.

Mouillaud Maurice et Têtu Jean-François (1989), *Le Jounal quotidien*, Lyon : Presses Universitaires de Lyon, 204p.

Mounier Evelyne (2013), Ch. 6 « L'information en milieu professionnel : le rôle des experts » in Clavier V. et Paganelli C. (dir.), *L'information professionnelle*, p. 129-150, Paris : Hermès Sciences Publications - Lavoisier, 2013.

Mounier Evelyne (1996), *Etude expérimentale de la segmentation d'un texte en paragraphes*. Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault et d'André Bisseret, Université Stendhal, Grenoble3, 1996.

Muller Charles (1973), *Initiation aux méthodes de la statistique linguistique,* Paris : Hachette, 1973, 187p.

Namer Fiammetta et Schmidt Paul (1997), « Construction d'un dictionnaire: morphologique à deux niveaux pour le français à l'aide de contraintes basées sur les structures de traits typées », *META : Journal des traducteurs*, 42 (1), p. 72-93.

Nie Jian-Yun (2003), « Introduction. Le domaine de la recherche d'information, survol d'une longue histoire », in Gaussier E. et Stéfanini M.-H. (dir.), *Assistance intelligente à la recherche d'informations*, p. 19-28, Paris : Hermès Science Publications – Lavoisier, 2003.

Nie Jian-Yun (1989), « A general information retrieval model based on modal logic », *Information Processing and Management*, 1989, 25 (5), p. 477-491.

Nie Jian-Yun et Savoy Jacques (2004), Ch. 3 « Modèles probabilistes en recherche d'informations. » In Ihadjadene M. (dir.), *Les systèmes de recherche d'informations : modèles conceptuels*, p. 55-76, Paris : Hermès Science Publications – Lavoisier, 2004.

Oger Claire et Ollivier-Yaniv Caroline (2007), « Analyse du discours et sociologie compréhensive : Retour critique sur une pratique de recherche transdisciplinaire », in Bonnafous S., Temmar M. (dir.), *Analyse du discours et sciences humaines et sociales,* p. 39-55, Paris : Ophrys, 2007. (Collection Les chemins du discours)

Olivesi Stéphane (dir.) (2006), *Sciences de l'information et de la communication. Objets savoirs, discipline.* Grenoble : PUG, 2006, 286p.

Ollivier Bruno (2007), *Les sciences de la communication. Théories et acquis*. Paris : Armand Colin, 2007, 284p.

Ollivier Bruno (2001), « « Enjeux de l'interdiscipline *», L'Année sociologique*, 2001, 51(2), p. 337-354.

Ouerfelli Tarek (2001), *La segmentation des documents techniques composites dans une perspective d'indexation. Vers la définition d'un modèle dans une optique d'automatisation.* Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, Université Stendhal, Grenoble3, 2001.

Paganelli Céline (2013), Ch. 10 « Les activités informationnelles en contexte de travail : questionnements en information-communication », p. 221-240, in Clavier V., Paganelli C. (dir.), *L'information professionnelle*, p. 221-240, Paris : Lavoisier Hermès Sciences Publications, 2013.

Paganelli Céline (2012), Une *approche info-communicationnelle des activités informationnelles en contexte de travail : acteurs, pratiques et logiques sociales,* Mémoire d'habilitation à diriger les recherches en sciences de l'information et de la communication, Université Stendhal, Grenoble 3, 2012.

Paganelli Céline, Mounier Evelyne et Pouchot Stéphanie (2011a), « Du papier au numérique : étude exploratoire des usages des collections de presse ancienne et des pratiques afférentes », in *Les intersections : gens, lieux, information, Actes du Congrès international ACSI/CAIS,* 2-4 juin 2011, Fredericton, 6p. [En ligne] URL : http://www.caisacsi.ca/proceedings/2011/28_Paganelli_Mounier_Pouchot.pdf Consultée le 19 janvier 2013.

Paganelli Céline, Mounier Evelyne et Pouchot Stéphanie (2011b) « Accès aux collections de presse ancienne : une étude exploratoire », in *Le "document" à l'ère de la différenciation numérique*, *Actes du 14ème Colloque International sur le document numérique*, p.249-266, Rabat, 7-9 décembre 2011, Editions Europia.

Paganelli Céline et Clavier Viviane (2011), « Le forum de discussion : une ressource informationnelle hybride entre information grand public et information spécialisée », p. 39-54. In Yasri-Labrique Eléonore (dir.) (2011), *Les forums de discussion : agoras du XXIe siècle? Théories, enjeux et pratiques discursives*, Paris : L'Harmattan, (coll. Langue et Parole).

Paganelli Céline (1997), *La recherche d'information dans des bases de documents techniques. Etude de l'activité des utilisateurs*, Thèse de doctorat en sciences de l'information et de la communication, sous la direction de Jacques Rouault, à l'Université Stendhal, Grenoble3, 1997.

Paijmans Hans (1993), « Comparing the Document Representations of two IR-Systems : CLARIT and TOPIC », Journal of the American Society for Information Science, 1993, 44 (7), p. 383-392.

Palmer Michael (2006), « Nommer les nouvelles du monde », *Hermès : Evénements mondiaux. Regards nationaux,* 2006 (46), p. 47-56.

Paveau Marie-Anne (2006), *Les prédiscours. Sens, mémoire, cognition,* Paris : Presses Sorbonne Nouvelle, 2006, 250p.

Pédauque Roger T. (2006), *Le document à la lumière du numérique*. *Forme texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité*, Caen : C&T éditions, 2006, 218p.

Pérennou Guy, Cotto Daniel, de Calmès Monique, Ferrané Isabelle et Pécatte Jean-Marie (1992), « *Le projet BDLEX de base de données lexicales du français écrit et parlé »* In *Actes du séminaire lexiques du GRECO-PRC Communication Homme-Machine,* p. 100-117, Toulouse, 21-22 Janvier 1992.

Péry-Woodley Marie Paule (2001), « Modes d'organisation et de signalisation dans des textes procéduraux », *Langages*, 2001, 35 (141), p. 28-46.

Péry-Woodley Marie-Paule (1995), « Quels corpus pour quels traitements automatiques? », *TAL*, 1995, 36(1-2), p. 213-232

Péry-Woodley Marie-Paule et Rebeyrolle Josette (1998), « Domain and genre in sublanguage text: definitional microtexts in three corpora », In Rubio A., Gallardo N., Castro R. & Tejada A. (ed.), *First International Conference on Language Resources and Evaluation,* p. 987-992, Paris: ELRA, 1998.

Péry-Woodley Marie-Paule and Scott Donia (2006), «Computational Approaches to Discourse and Document Processing », *T.A.L.,* 2006, 47(2), 2006, p 7-19.

Piotrowski David (1999), « Le T.A.L.N. Faillite des ambitions théoriques, Succès de la raison pratique », *Sémiotiques,* 1999 (17), p. 1-27.

Polity Yolla, Henneron Gérard, Palermiti Rosalba (ed.) (2005), *L'organisation des connaissances. Approches conceptuelles*, Actes du 4è congrès ISKO-France, 3 et 4 juillet 2003, Paris, Budapest, Turin : L'Harmattan, I2005, 266p. (La librairie des humanités)

Ponton Claude (1996), *Génération automatique de textes en langue naturelle : essai de définition d'un système noyau*. Thèse de doctorat en sciences de l'information et de la communication sous la direction de Jacques Rouault, Université Stendhal, Grenoble3, 1996.

Porter Martin F. (1980), « Algorithm for Suffix Stripping », *Program,* 1980, 14 (3), p. 130-137.

Poudat Céline, Cleuziou Guillaume et Clavier Viviane (2006), « Catégorisation de textes en domaines et genres : complémentarité des indexations lexicale et morphosyntaxique », *Document numérique*, vol. 9, n°1/2006, Lavoisier - Hermès, p. 61-76.

Poudat Céline (2006), *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, thèse de doctorat en linguistique, sous la direction de Gabriel Bergounioux, Université d'Orléans, 2006. [En ligne] URL : http://refef.crifpe.ca/document/these/POUDAT.pdf Consultée le 16 mai 2013.

Rabatel Alain (2004), « L'effacement énonciatif dans les discours rapportés et ses effets pragmatiques »*., Langages*, 2004, 38 (156), p. 3-17.

Rastier François (2005), « Pour une sémantique des textes théoriques », *Revue de sémantique et de pragmatique,* 2005 (17), p. 151-180. [En ligne sur *Texto !*] URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Textes.html Consultée le 23 avril 2013.

Rastier François (2004a), « Enjeux épistémologiques de la linguistique de corpus » *Texto !*[En ligne] URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html Consultée le 20 avril 2013.

Rastier François (2004b), « Ontologie(s) », *Revue des sciences et technologies de l'information,* 2004, 18 (1), p. 15-40. [En ligne sur *Texto !*] URL : www.revue-texto.net/Inedits/Rastier/Rastier_Ontologies.html Consultée le 20 avril 2013.

Rastier François (1998), « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, 1998, 32 (129), p. 97-111.

Rastier François (1996), « *La* sémantique des thèmes ou le voyage sentimental », *Texto !* [En ligne]. URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html Consultée le 25 mai 2013.

Rastier François (1989), *Sens et textualité,* Paris : Hachette, 286p. (Langue, linguistique, communication)

Rebeyrolle Josette et Péry-Woodley Marie-Paule (1998), « Repérage d'objets textuels fonctionnels pour le filtrage d'information : le cas de la définition », Rencontre

Internationale sur l'Extraction "*Le Filtrage et le Résumé Automatique"*, RIFRA'98, ARIS/CNRS, p. 19-30.

Reneker Maxine H. (1993), « A qualitative study of information seeking among members of an academic community:methodological issues and problems », *The Library Quarterly*, 1993, 63(4), p. 487-507.

Rinck Fanny, Boch Françoise et Grossmann Francis (2007), « Quelques lieux de variation du positionnement énonciatif dans l'article de recherche », in Lambert P., Millet A., Rispail M. et Trimaille C. (dir.), *Variations au cœur et aux marges de la sociolinguistique – Mélanges offerts à Jacqueline Billiez,* p. 285-296, Paris : L'Harmattan, 2007. (Espaces Discursifs)

Ringoot Roselyne et Robert-Demontrond Philippe (2004), *L'analyse de discours*, Editions Apogée, 2004, 222 p.

Robertson Stephen E. (1977), « The probabilité ranking principle in IR », *Journal of Documentation*, 1977, 33 (4), p. 294-304.

Rouault Jacques (1987), *La linguistique automatique. Applications documentaires.* Berne, Francfort/Main, Paris : Peter Lang, 1987, 309p.

Rouault Jacques (1971), *Approche formelle de problèmes liés à la sémantique des langues naturelles*, thèse de doctorat en mathématiques sous la direction de Bernard Vauquois, Université Joseph Fourrier, 1971.

Rouault Jacques et Miège Bernard (2003), Ch.1 « Approches et fondements », in Miège B. (dir.), *Communication personnes systèmes informationnels*, p 21-40, Paris : Hermès Science Publications – Lavoisier, 2003.

Salaün Jean-Michel (1993), « Les sciences de l'information en question. Le point de vue du lecteur », *Réseaux,* 1993, 11 (58), n° 58, p. 9-25.

Salton Gerard (1991), *The state of Retrieval System Evaluation*, Computer science technical report, Cornell University, May 1991, 19p. [En ligne]
URL : http://hdl.handle.net/1813/7046 Consultée le 11 septembre 2013.

Salton Gerard (ed.) (1971), *The SMART Retrieval System – Experiments in Automatic Document Processing*, Prentice Hall, New-Jersey : Englewood Cliffs, 1971.

Salton Gerard (1968), *Automatic Information Organization and Retrieval*, New-York : McGraw-Hill, 1968.

Saracevic Tefko (1997), « The stratified model of information retrieval interaction : Extension and applications », *In Proceedings of the American Society for Information Science*, volume 34, pages 313–327.

Savolainen Reijo (2008), *Everyday information practices: A social phenomenological perspective,* Lanham : Scarecrow Press 2008, 233p.

Schatzki Theodore, Knorr-Cetina Karin, Savigny Eike von (ed.) (2001), *The practice turn in contemporary theory,* London: Routledge, 2001, 256p.

Simonnot Brigitte (2012), *L'accès à l'information en ligne. Moteurs, dispositifs et médiations*, Paris : Hermès-Lavoisier, 256p. (collection Systèmes d'information et organisation documentaire)

Sparck-Jones Karen (1974), « Automatic Indexing », *Journal of Documentation,* December 1974, 30 (4), p. 393-429.

Sparck-Jones Karen (1992), « Information retrieval », in Stuart C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence,* p. 684-690, 1992, I-II, New York : John Wiley & Sons.

Sparck-Jones Karen (1999), « What is the role of NLP in Information Retrieval ? », in Strzalkowski T. (ed.), *Natural Language Information Retrieval,* p. 1-24, Dordrecht : Kluwer Academie Publishers.

Staii Adrian (2012), *Grammaires sociotechniques des Tics numériques. Pour une théorie élargie de l'ancrage social.* Mémoire d'habilitation à diriger les recherches en sciences de l'information et de la communication, Université Stendhal, Grenoble3, 2012.

Staii Adrian, Balicco Laurence, Bertier Marc, Clavier Viviane, Mounier Evelyne et Paganelli Céline (2008), « Les pratiques informationnelles des médecins dans les centres hospitaliers universitaires : au croisement de la logique scientifique et de la culture professionnelle », Revue canadienne des sciences de l'information et de bibliothéconomie, vol. 30, n°1/2, mars-juin 2006, p. 69-90.

Staii Adrian (2004), « Réflexion sur les recherches et le champ des sciences de l'information », *Les Enjeux de l'information et de la communication,* 15p. [En ligne] URL : http://w3.u-grenoble3.fr/les_enjeux/2004/Staii/staii.pdf Consultée le 14 juillet 2013.

Talia Sanna, Keso Heidi, Pietiläinen Tarja (1999), « The production of `context' in information seekingresearch: a metatheoretical view », *Information Processing & Management,* 1999, 35 (6), p. 751-763.

Tamine Lynda et Calabretto Sylvie (2008), Ch. 7 « Recherche d'information contextuelle et web. », In Boughanem M. et Savoy J. (ed.), *Recherche d'information. Etat des lieux et perspectives*, p. 201-230, Paris : Hermès Science Publications - Lavoisier 2008.

Tanguy Ludovic et Hathout Nabil (2002), « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web », *9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles*, *TALN 2002*, Nancy, 24–27 juin 2002. [En ligne] URL : http://redac.univ-tlse2.fr/lexiques/verbaction/Hathout-2002-TALN.pdf Consultée le 12 mars 2013.

Teissier Marc (2010) *La numérisation du patrimoine écrit*, Janvier 2010, Rapport, *La documentation française,* 64p. [En ligne]

URL : http://www.ladocumentationfrancaise.fr/rapports-publics/104000016/index.shtml Consulté le 19 mars 2013.

Tennis Joseph T. (2013), « Le poids du langage et de l'action dans l'organisation des connaissances : position épistémologique, action méthodologique et perspective théorique », *Études de communication*, 2013 (39), p. 41-64.

Timimi Ismaïl (1999), *De la paraphrase linguistique à la recherche d'information, le système 3 AD : théorie et implantation (aide à l'analyse automatique du discours)*, Thèse de doctorat en informatique sous la direction de Jacques Rouault, Université Grenoble 3, 1999.

Timimi Ismaïl (1998), « Gouverner par le sondage & analyser par la paraphrase. Méthode TIAD », *Actes des JADT 1998*, 615p. [En ligne] URL : http://lexicometrica.univ-paris3.fr/jadt/jadt1998/timimi.htm Consultée le 7 février 2013.

Timimi Ismaïl et Rouault Jacques (1997), « La paraphrase comme relation d'équivalence dans l'analyse automatique du discours », *Actes de TALN'97,* [En ligne] URL : http://www.lirmm.fr/~lafourcade/ML-pool/Mathieu_Lafourcade/mathieu_lafourcade-68.html Consultée le 7 février 2013.

Tutin Agnès (2010), *Sens et combinatoire lexicale : de la langue au discours,* Mémoire d'habilitation à diriger les recherches en sciences du langage, Volume 1, Université Stendhal, Grenoble 3, 2010. [En ligne] URL : http://w3.u-grenoble3.fr/lidilem/labo/file/HDR_Tutin.pdf Consultée le 12 janvier 2013.

Tutin Agnès (2007), « Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques », *Actes de TALN'07*, Toulouse, 5-8 juin 2007. [En ligne]URL : http://w3.u-grenoble3.fr/tutin/taln_tutin_final_avril.pdf Consultée le 12 janvier 2013.

Vakkari, P. (1997), « Information seeking in context: a challenging metatheory », In P. Vakkari P., Savolainen R. & Dervin B., *Information seeking in context, in*Proceedings of an International Conference on Research in Information Needs, Seeking and use in Different Contexts, pp. 451-464 London: Taylor Graham.

Valette Mathieu et Slodzian Monique (2008), « Sémantique des textes et recherche d'information », *Revue française de linguistique appliquée,* 2008, XIII (1), p. 119-133.

Van Rijsbergen C.J. (1979), *Information Retrieval*. 2nd édition. London, UK : Butterworths, 1979.
[En ligne] URL : http://www.dcs.gla.ac.uk/Keith/Preface.html Consultée le 16 décembre 2012.

Van Rijsbergen, C.J. (1986), « A New Theoretical Framework for Information Retrieval », *ACM Conference on Research and Development in Information Retrieval,* Pise, Italie, 1986, p. 194-200.

Vold Thue Eva (2008), *Modalité épistémique et discours scientifique. Une étude contrastive des modalisateurs épistémiques dans des articles de recherche français, norvégiens et anglais, en linguistique et médecine.* Thèse de doctorat en philosophie, Université de Bergen, 2008. [En ligne]
URL : https://bora.uib.no/bitstream/1956/2653/1/Dr.Avh._Eva_T_Vold.pdf
Consultée le 25 juin 2013.

Voorhees Ellen M. (1993), « Using *WordNet* to disambiguiate word senses for text retrieval », Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in information retrieval, SIGIR'93, New-York : ACM, p. 171-180.

Warren Karine (1998), *Gestion de conflits dans une architecture multi-agents d'analyse automatique de texte*, thèse de doctorat en sciences de l'information et de la communication sous la direction de Jacques Rouault, Université Grenoble3, 1998.

Weick Karl E. (2001), *Making sense of the organization*, Oxford : Blackwell Publishers, 2001, 483 p.

Weick Karl E. (1995), *Sensemaking in organization*, Thousand Oaks : Sage Publications, 1995, 231p.

Westeel Isabelle (2009), « Le patrimoine passe au numérique », *BBF,* 2009, 54 (1), p. 28-35.

Wilson Tom D. (2010), « Fifty Years of Information Behavior Research », *American Society for Information Science and Technology,* February-March 2010, 36 (3), p. 27-34. [En ligne] URL : http://www.asis.org/Bulletin/Feb-10/FebMar10_Wilson.html Consulté le 2 avril 2013.

Wilson Tom D. (2002), « Schutz, phenomenology and research methodology for information behaviour research », *Paper presented at the Fourth International Conference on Information Seeking in Context*, Lisbon, September. [En ligne]
URL : http://comminfo.rutgers.edu/~belkin/612-05/wilson-schutz.pdf
Consulté le 2 avril 2013.

Wilson Tom D. (1999), « Models in information behaviour research », *Journal of Documentation*, 1999, 55 (3), p. 249-70.

Wilson Tom D. (1997), « Information behaviour: an interdisciplinary perspective », *Information Processing and Management,* 1997, 33 (4), p. 551-572.

Wilson Tom D. (1994), « Information needs and uses: 50 years of progress? »,, in Vickery, B.C. (Ed.), *Fifty Years of Information Progress: A Journal of Documentation Review*, Aslib, London, p. 15-51. [En ligne]
URL : http://informationr.net/tdw/publ/papers/1994FiftyYears.html
Consultée le 12 juillet 2013.

Wilson Tom D. (1981), « On user studies and information needs », *Journal of Documentation*, 1981, 37 (1), p. 3-15.

Yahiaoui Leila, Prié Yannick et Boufaida Zizette (2008), « Redocumentation des traces d'activité médiée informatiquement dans le cadre des transactions communication », *XIXe Journées francophones d'ingénierie des Connaissances, IC'2008*, 18-20 juin 2008, Nancy, p. 1-13. [En ligne] URL : http://liris.cnrs.fr/Documents/Liris-3446.pdf Consultée le 16 juillet 2013.

Zacklad Manuel (2010), « Évaluation des systèmes d'organisation des connaissances », *Les Cahiers du numérique*, 2010, 6 (3), p. 133-166.

Zacklad Manuel (2007), « Classification, thésaurus, ontologies, folksonomies : comparaisons du point de vue de la recherche ouverte d'information (ROI) », *35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté : Franchir les frontières,* 2007. [En ligne]URL : http://hal.archives-ouvertes.fr/docs/00/20/24/40/PDF/cais-acsi_zacklad_-_avec_ref.pdf Consultée le 16 juin 2013

Zobel Justin and Moffat Alistair (2006), « Inverted Files for Text Search Engines», *ACM Computing Surveys*, 2006, 38(2), p. 1-56.

Annexe 1 : Extrait du corpus CFM balisé suivant la norme TEI

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <TEI.2>
- <teiHeader type="text">
- <fileDesc>
- <titleStmt>
<title>Version numérique de 0002 Moteur thermoacoustique annulaire à ondes progressives</title>
- <respStmt>
<resp>collecté et balisé par</resp>
<name>Viviane Clavier</name>
</respStmt>
</titleStmt>
- <publicationStmt>
<distributor>Coral</distributor>
</publicationStmt>
- <sourceDesc>
<author>Stéphane JOB, Vitalyi GUSEV1, Pierrick LOTTON, Michel BRUNEAU.</author>
<docDate>value = "2001"</docDate>
</sourceDesc>
</fileDesc>
- <encodingDesc>
- <editorialDecl>
- <correction>
<p>Aucune correction</p>
</correction>
</editorialDecl>
</encodingDesc>
- <profileDesc>
- <langUsage>
<language id="FR" />
</langUsage>
- <textClass>
<catRef target="mécanique accoustique" scheme="discipline" />
<catRef target="article" scheme="genre" />
<catRef target="expert" scheme="auteur" />
</textClass>
</profileDesc>
<revisionDesc />
</teiHeader>
- <text>
- <front>
<docTitle>Moteur thermoacoustique annulaire à ondes progressives.</docTitle>
<docAuthor>Stéphane JOB, Vitalyi GUSEV1, Pierrick LOTTON, Michel BRUNEAU.</docAuthor>
<affAuthor>Laboratoire d'Acoustique, UMR CNRS 6613. 1 Laboratoire de Physique de l'Etat Condensé, UMR CNRS 6087, ENSIM. Université du Maine, Avenue Olivier Messiaen, 72085 LE MANS CEDEX 9 - France</affAuthor>
- <div type="résumé">
<head>Résumé</head>
<p>Les moteurs thermoacoustiques font l'objet d'un regain d'intérêt en raison des applications potentielles qui se dessinent de nos jours. L'objet de cette communication est de proposer quelques-uns des résultats récents concernant les moteurs thermoacoustiques annulaires à ondes progressives. Plus particulièrement, l'étude du régime transitoire est proposée (déclenchement de l'instabilité thermoacoustique), ainsi que celle des processus de stabilisation, conduisant au régime permanent (processus de cascade harmonique, génération d'un écoulement massique stationnaire induit, modification du comportement thermique du système).</p>
</div>
- <div type="abstract">
<head>Abstract</head>
<p>Thermoacoustic prime-movers are the subject of a renewed interest due to potential applications. The goal of this presentation deals with some of the recent results concerning travelling wave annular thermoacoustic prime-movers. Particularly, the study of transient regime is proposed (thermoacoustic instability release), as well as the stabilisation processes, leading to a stationary regime (harmonic cascade process, generation of an induced stationary mass flow, modification of the thermal behaviour of the system).</p>
</div>
- <div type="mots-clés">
<head>Mots-clés</head>
<p>Acoustique, thermoacoustique, phénomènes non linéaires, instabilité.</p>
</div>
```

</front>
- <body>
-
<head>1. Introduction.</head>
-
Depuis deux décennies, de nombreux travaux, initiés par N. Rott (1980), relevant de la thermoacoustique, portent sur l'étude des phénomènes d'interactions qui existent entre énergie acoustique et énergie thermique au sein des couches limites acoustiques, au voisinage de parois. Parmi les applications possibles de ces phénomènes, on trouve (Wheatley et col. (1985), Swift (1988)) des pompes à chaleur (ou réfrigérateur) et des moteurs, que l'on peut eux-mêmes distinguer en deux catégories suivant qu'ils fonctionnent en ondes stationnaires ou progressives. Le système à l'étude, représenté sur la figure 1, est un moteur thermoacoustique annulaire à ondes progressives. Il est constitué d'un guide d'onde de section cylindrique (de diamètre
<formula>53mm</formula>
) et de longueur
<formula>L=2m12</formula>
) rempli d'un fluide gazeux (de l'air à pression ambiante), de forme toroïdal : la propagation acoustique est en onde plane et le trajet acoustique est bouclé sur lui-même. Un élément actif appelé noyau thermoacoustique est disposé dans ce guide d'onde. Il est constitué d'un empilement de parois parallèles à l'axe du tube et espacées d'une distance de quelques épaisseurs de couche limite acoustique. Un gradient de température est maintenu dans le noyau thermoacoustique au moyen d'échangeurs thermiques. L'effet thermoacoustique étant un effet pariétal, le noyau thermoacoustique est destiné à augmenter la surface de contact entre le fluide et les parois. Dans le prototype expérimental, un certain nombre de thermocouples permettent de mesurer cette distribution de température imposée, et deux microphones permettent de mesurer le champ de pression acoustique et de séparer ce dernier en ses deux contributions contra-propagatives. Une analyse de Fourier par détection synchrone permet encore d'avoir accès au contenu spectral de l'onde générée dans le système.

- <figure type="schéma">
<head>FIG. 1 : Présentation schématique d'un moteur thermoacoustique annulaire</head>
</figure>
Dans le paragraphe 2, l'étude du régime transitoire est abordée en analysant les modalités de déclenchement de l'instabilité thermoacoustique et de génération d'une onde acoustique progressive. L'obtention du régime stationnaire de fonctionnement du moteur annulaire, traité dans le paragraphe 3, est distinguée sous deux aspects. La formation d'une onde de choc, en regard des fort niveaux sonores générés, et le processus de saturation non linéaire de l'amplitude acoustique, sont étudiés dans le paragraphe 3.1. D'autres phénomènes induits, pouvant expliquer les résultats de l'observation du régime permanent, tels que la génération d'un écoulement stationnaire du fluide ou l'accroissement de la conductivité du gaz, sont donnés dans le paragraphe 3.2.

-
<head>2. Régime transitoire : déclenchement de l'instabilité thermoacoustique.</head>
-
Il possible de montrer que le noyau thermoacoustique, tel qu'il est présenté sur la figure 1 (
<formula>x ([,0 H s + H]w</formula>
), dans lequel le gradient de température est imposé, est un élément actif capable d'amplifier une onde acoustique qui le traverse dans une direction donnée (de la gauche vers la droite sur la figure 1), tandis que l'onde se propageant dans l'autre direction est atténuée. Une partie du flux d'énergie thermique fournie pour maintenir le gradient de température est utilisée pour augmenter l'énergie de l'onde acoustique entrante. Lorsque cette amplification compense exactement les pertes thermo-visqueuses au cours de la propagation acoustique dans la partie passive du guide d'onde (
<formula>x ~ [H S + H W ,]L</formula>
), sur la figure 1), le système devient instable et la présence d'une quelconque excitation peut ainsi déclencher l'instabilité thermoacoustique, et donner naissance à un phénomène oscillant.

- <figure type="schéma">
<head>FIG. 2 : Déclenchement de l'instabilité thermoacoustique : amplitude de pression mesurée (trait plein) et estimé à partir des pertes par cascade harmonique (trait pointillé).</head>
</figure>
-
La géométrie annulaire du guide d'onde favorise la génération d'une onde progressive, dont l'amplitude augmente à chaque passage dans l'élément actif. L'amplitude de la pression acoustique augmente alors de façon exponentielle, au cours des premières secondes qui suivent le déclenchement de l'instabilité. L'amplitude, notée
<formula>+ 1H</formula>
, de l'onde se propageant dans le sens des x croissants à la fréquence fondamentale
<formula>+ (f ~ 152 Hz</formula>
), normalisée à la pression atmosphérique
<formula>+ 2 00c~</formula>
, , est représenté en trait plein sur la figure 2. On note qu'au cours des dix premières secondes, le régime transitoire se caractérise par une croissance exponentielle de
<formula>+ 1H</formula>

<head>3. Régime stationnaire : phénomènes non linéaires.</head>
Le phénomène transitoire étant pleinement déclenché, il est alors nécessaire de prendre en compte la présence de phénomènes limitants pour décrire l'obtention d'un régime stationnaire et permanent. En regard des forts niveaux sonores générés par ce type de système, ces facteurs limitants sont de type non linéaire. Dans la suite, on distingue l'effet classique provenant de la propagation acoustique non linéaire à fort niveaux (saturation de l'amplitude), de la génération de phénomènes non linéaires induits particuliers, tels que la génération d'un écoulement stationnaire dans le fluide (à l'origine de la convection d'une partie de l'énergie thermique), ou encore l'accroissement de la conductivité thermique du fluide (accroissement du flux thermique par conduction). Ces deux phénomènes modifient les caractéristiques thermiques du système et par voie de conséquence son aptitude à amplifier l'onde acoustique.

<head>3.1. Onde de choc et cascade harmonique.</head>
Le premier phénomène éligible permettant d'expliquer l'obtention du niveau stationnaire de l'amplitude de l'onde acoustique généré est le début de la formation d'une onde de choc (Job et col., (2000)). En effet, la propagation acoustique à forts niveaux sonores s'accompagne de la génération d'harmoniques supérieurs (Atchley et col. (1990), Swift (1992)) dans le contenu spectral de l'onde de pression. Une partie de l'énergie acoustique oscillant à la fréquence fondamentale est transférée, par couplage non linéaire, vers les hautes fréquences : c'est le processus de cascade harmonique. Ce phénomène se traduit par la saturation de l'amplitude totale de l'oscillation, et par la modification du profil temporel de cette onde (depuis une sinusoïde, à faible amplitude, vers un profil contenant une onde de choc, à forte amplitude). Les observations accréditent l'hypothèse d'un comportement faiblement non linéaire. L'évolution du contenu spectral de la pression acoustique mesurée est représenté sur la courbe de gauche de la figure 3 : l'importance relative des deux premiers harmoniques par rapport au fondamental augmente quand le niveaux sonore augmente. Sur la courbe de droite de la figure 3, la reconstruction du profil temporel de l'onde sur une période d'oscillation, pour la température de chauffage la plus importante, permet de visualiser le début de la formation d'une onde de choc (la courbe en pointillés est une sinusoïde).
<figure type="schéma">
<head>
FIG. 3 : Mesure des trois premières composantes harmoniques
<formula />
de la pression acoustique se propageant dans le sens des x croissants, en fonction de la température du point chaud (courbe de gauche). Reconstruction du profil temporel de cette onde (courbe de droite).
</head>
</figure>
Cependant, la prise en compte de la formation d'une onde de choc n'est pas suffisante pour prédire le niveau de pression acoustique du régime stationnaire, comme le montre la courbe en pointillés de la figure 2. Cette courbe est une estimation du niveau du fondamental qui serait observé par le seul effet de cascade harmonique. Or cette estimation est encore trop importante, comparée au niveau effectivement mesuré, et il convient alors d'étudier la présence d'autres phénomènes non linéaires susceptibles de limiter l'amplitude en régime permanent.

<head>3.2. Ecoulement redressé et conductivité induite.</head>
En présence d'une onde acoustique de forte amplitude, il est possible de prévoir la génération d'un écoulement stationnaire de fluide, par effets non linéaires (Gusev et col. (2000)). Cet écoulement va d'une part engendrer un phénomène de convection dans le bilan énergétique global, et d'autre part modifier progressivement la distribution de température dans le noyau thermoacoustique. Il en est de même pour un autre phénomène, qui résulte du mouvement des particules de fluide sous l'effet de l'onde acoustique : ce déplacement induit alors une augmentation de la conductivité du fluide, dont l'effet est identique au précédent.
<figure type="schéma">
<head>FIG. 4 : Distribution de température dans le noyau thermoacoustique, en dessous (traits pointillés) et au dessus (traits pleins) du seuil de déclenchement.</head>
</figure>

Une mesure de la distribution de température normalisée C est donnée sur la figure 4, où
<formula>HT et CT</formula>
sont les températures des points chauds et froids et
<formula>mT</formula>
est la température moyenne du fluide : les différences observées sur la forme de la distribution de température peuvent alors être corrélées à la présence de ces deux phénomènes.

<figure type="schéma">
<head>FIG. 5 : Estimations théoriques et expérimentales de la conductivité induite (courbe de gauche) et de la vitesse de l'écoulement redressé (courbe de droite).</head>
</figure>

En décomposant le flux d'enthalpie en tout point du noyau thermoacoustique, il est possible de relier la distribution de température à la présence d'un écoulement massique et d'une conductivité accrue du fluide. Les résultats de l'analyse de la mesure de température en différents points du noyau thermoacoustique (Job et col. (2000)) sont présentés sur la figure 5 : la courbe de gauche est une estimation de la conductivité induite

<formula>ack</formula>

dans l'empilement de parois, et la courbe de droite est une estimation de la vitesse moyenne

<formula>xmv</formula>

de l'écoulement massique du fluide dans le guide d'onde. Les courbes en trait plein sont des prédictions théoriques (Gusev et col. (2000)). L'état actuel de nos travaux ne permet pas pour l'instant de quantifier la dépendance du niveau d'amplification dans le noyau thermoacoustique en fonction de la distribution de température, mais seulement d'affirmer que cette dépendance existe. Ces travaux, en cours d'investigation, visent entre autre à obtenir une formulation analytique du phénomène d'amplification thermoacoustique, pour une distribution quelconque de température. Pour l'instant, la simple comparaison du temps caractéristique d'établissement du régime permanent et de ceux liés aux phénomènes étudiés tend à faire penser que ces phénomènes entrent en jeu lors de la stabilisation du système.

-
<head>4. Conclusions.</head>
L'étude des moteurs thermoacoustiques annulaires a été abordée en s'intéressant d'abord au régime transitoire d'établissement de l'instabilité thermoacoustique, puis aux phénomènes de stabilisation du régime permanent. La seule prise en compte du phénomène classique de cascade harmonique n'est pas suffisante pour expliquer l'obtention du niveau sonore en régime permanent. Cette stabilisation s'explique aussi par la présence de phénomènes induits qui agissent sur la distribution de température, origine de l'amplification thermoacoustique. Ce travail bénéficie du soutien de la DGA (contrat 99-34-072/DSP) et d'une bourse DGA-CNRS pour le premier auteur.

</body>
- <back>
-
<head>Références</head>
Rott, N. , 1980 Thermoacoustics, Adv. Appl. Mech., 20, pp. 135-175. Wheatley, J., Hofler, T., Swift, G.W., Migliori, A., 1985 Understanding some simple phenomena in thermoacoustic with applications to acoustical engines, Am. J. Phys., 53(2), pp. 147-162. Swift, G.W., 1988 Thermoacoustic engines, J. Acoust. Soc. Am., 84(4), pp. 1145-1180.Job, S., Gusev, V., Lotton, P., Bruneau, M., 2000, Weakly nonlinear acoustic wave in a travelling wave thermoacoustic engine, C. R. Acad. Sc., submitted. Atchley, A.A., Bass, H.E., Hofler, T.J., 1990, Development of nonlinear waves in a thermoacoustic prime-mover, in Frontiers of Nonlinear Acoustics 12th ISNA, Elsevier, New York, pp.603-608.Swift, G.W., 1992, Analysis and performance of a large thermoacoustic engine, J. Acoust. Soc. Am., 92, pp. 1551-1563.Gusev, V., Job, S., Bailliet, H., Lotton, P., Bruneau, M., 2000, Acoustic streaming in an annular thermoacoustic prime-mover, J. Acoust. Soc. Am., 108(3), pp. 934-945. Job, S., Gusev, V., Lotton, P., Bruneau, M., 2000, On the velocity of the acoustic streaming in annular thermoacoustic prime-movers, Appl. Phys. Lett., submitted.

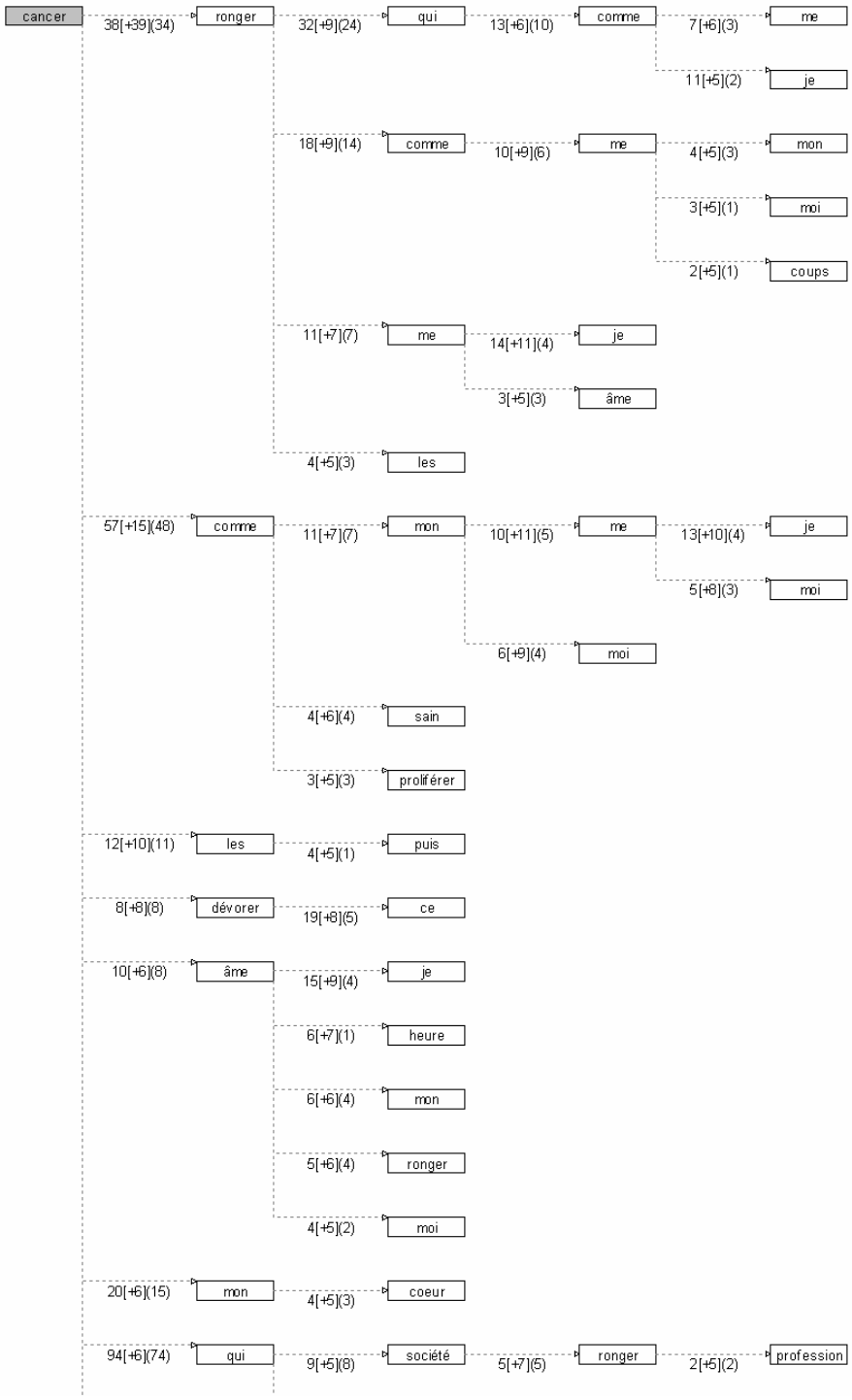
</back>
</text>
</TEI.2>

Annexe 2 : Extrait du corpus CFM étiqueté et post-édité

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <teiCorpus>
- <text ana="323">
  <w ana="NC:sg">moteur</w>
  <w ana="ADJ:sg">thermoacoustique</w>
  <w ana="ADJ:sg">annulaire</w>
  <w ana="PREP">à</w>
  <w ana="NC:pl">ondes</w>
  <w ana="ADJ:pl">progressives</w>
  <w ana="PON:dot">.</w>
  <w ana="NP">Stéphane</w>
  <w ana="NP">JOB</w>
  <w ana="PON:comma">,</w>
  <w ana="NP">Vitalyi</w>
  <w ana="NP">GUSEV1</w>
  <w ana="PON:comma">,</w>
  <w ana="NP">Pierrick</w>
  <w ana="NP">LOTTON</w>
  <w ana="PON:comma">,</w>
  <w ana="NP">Michel</w>
  <w ana="NP">BRUNEAU</w>
  <w ana="PON:dot">.</w>
  <w ana="NC:sg">Laboratoire</w>
  <w ana="PREP">d'</w>
  <w ana="NC:sg">Acoustique</w>
  <w ana="PON:comma">,</w>
  <w ana="SIG">UMR</w>
  <w ana="SIG">CNRS</w>
  <w ana="NUM:post">6613.1</w>
  <w ana="NC:sg">Laboratoire</w>
  <w ana="PREP">de</w>
  <w ana="NP">Physique</w>
  <w ana="PREP">de</w>
  <w ana="DET:def">l'</w>
  <w ana="NP">Etat</w>
  <w ana="NP">Condensé</w>
  <w ana="PON:comma">,</w>
  <w ana="SIG">UMR</w>
  <w ana="SIG">CNRS</w>
  <w ana="NUM:post">6087</w>
  <w ana="PON:comma">,</w>
  <w ana="SIG">ENSIM</w>
  <w ana="PON:dot">.</w>
  <w ana="NC:sg">Université</w>
  <w ana="DTC:sg">du</w>
  <w ana="NP">Maine</w>
  <w ana="PON:comma">,</w>
  <w ana="NC:sg">Avenue</w>
  <w ana="NP">Olivier</w>
  <w ana="NP">Messiaen</w>
  <w ana="PON:comma">,</w>
  <w ana="NUM:post">72085</w>
  <w ana="DET:def">LE</w>
  <w ana="PREP">MANS</w>
  <w ana="NC:sg">CEDEX</w>
  <w ana="NUM:post">9</w>
  <w ana="PON:tiret">-</w>
  <w ana="DET:dem">FRANCE</w>
  <w ana="NC:sg">Résumé</w>
  <w ana="PON:colon">:</w>
  <w ana="DET:def">Les</w>
  <w ana="NC:pl">moteurs</w>
  <w ana="ADJ:pl">thermoacoustiques</w>
  <w ana="VER:pres">font</w>
  <w ana="DET:def">l'</w>
  <w ana="NC:sg">objet</w>
  <w ana="PREP">d'</w>
  <w ana="DET:indef">un</w>
  <w ana="NC:sg">regain</w>
  <w ana="PREP">d'</w>
  <w ana="NC:sg">intérêt</w>
  <w ana="PREP">en</w>
  <w ana="NC:sg">raison</w>
  <w ana="DTC:pl">des</w>
  <w ana="NC:pl">applications</w>
  <w ana="ADJ:pl">potentielles</w>
  <w ana="PRO:rel">qui</w>
  <w ana="PRO:refl">se</w>
  <w ana="VER:pres">dessinent</w>
  <w ana="PREP">de</w>
  <w ana="DET:poss:pp1pl">nos</w>
  <w ana="NC:pl">jours</w>
  <w ana="PON:dot">.</w>
  <w ana="DET:def">L'</w>
  <w ana="NC:sg">objet</w>
  <w ana="PREP">de</w>
  <w ana="DET:dem">cette</w>
  <w ana="NC:sg">communication</w>
  <w ana="VER:pres">est</w>
  <w ana="PREP">de</w>
  <w ana="VER:inf">proposer</w>
  <w ana="DET:indef">quelques-</w>
  <w ana="PRO:indef">uns</w>
  <w ana="DTC:pl">des</w>
  <w ana="NC:pl">résultats</w>
  <w ana="ADJ:pl">récents</w>
  <w ana="VER:parpres">concernant</w>
  <w ana="DET:def">les</w>
  <w ana="NC:pl">moteurs</w>
```

Annexe 3 :Graphe de cooccurrences des valeurs figurées du mot *cancer* dans les corpus *Frantext*, *Le Figaro*, et le *Nouvel Obs*.

Le graphe est produit par le logiciel Coocs et les trois séries de nombres qui apparaissent désignent respectivement le seuil de fréquence de la forme coocurrente dans le corpus, le seuil de co-fréquence et le seuil de spécificité.



Annexe 4 : Graphe de cooccurrence des valeurs propres du mot *cancer* dans les corpus *Frantext et le NouvelObs*.

Le graphe est produit par le logiciel Coocs et les trois séries de nombres qui apparaissent désignent respectivement le seuil de fréquence de la forme coocurrente dans le corpus, le seuil de co-fréquence et le seuil de spécificité.